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

Joy Olusanya[1]
Masakhane NLP
Obafemi Awolowo University
joyolusanya@student.oauife.edu.ng

Daud Abolade[1]
Masakhane NLP
University of Lagos
aboladedawud@gmail.com

Abstract

Natural Language Processing (NLP) for African languages such as Yoruba remains underdeveloped due to limited annotated resources, linguistic variability, and a lack of specialized models. In this paper, we present ÒWE-YOR, a Yoruba proverb dataset that considers text classification. The fact that Yoruba proverbs are an important ingredient of Yoruba cultural heritage and also day-to-day interaction presses the need for NLP models that are sensitive and inclusive of linguistic diversity. Our work leverages a balanced dataset of 15,925 labeled entries, out of which are 7,963 proverbs and 7,962 non-proverbs, carefully collected and annotated. Our study proposes a machine learning and a transformer-based methodology that involves training Naive Bayes Algorithm, then fine-tuning existing language models like BERT multilingual case and AfroLM to learn the contextual features specific to Yoruba proverbs.

1 Introduction

In the world of Natural Language Processing today, text classification and text understanding have grown to a level where a series of language models perform well on these tasks in various high-resource languages, but there is still work to be done for low-resource languages in such tasks, especially for culturally rich languages like Yoruba. Human language is not just a tool for communication but also a vault of culture, identity, and shared wisdom. In Yoruba culture, the proverb is an important tool that encodes cultural wisdom and requires the ability to decode before it can be understood. It requires a high level of reasoning and understanding of indigenous knowledge for machines to produce high-level results. Bringing up models capable of classifying text is a major challenge, especially for under-resourced languages. With the aim of improving the text classification dataset, there have been various works toward the curation of dataset (Rassi et al., 2014; Pighin et al., 2019). However, most of these works do not take into consideration proverbs in other African languages.

Existing NLP research on African languages has focused on a series of efforts in developing datasets for various NLP tasks such as news topic classification (Ifeoluwa Adelani et al., 2023; Niyongabo et al., 2020), named entity recognition (Adelani et al., 2022b), machine translation (Adelani et al., 2022a; Costa-jussà et al., 2022), and part-of-speech tagging (Nivre et al., 2016). However, few have attempted to study culturally rich datasets such as proverbs for classification tasks. To overcome this, we present a Yoruba proverb dataset, ÒWE-YOR, which supports the classification of proverbs and non-proverbs.

Our work evaluates machine learning and transformer-based models for Yoruba proverb classification. We trained a MultinomialNB model and later fine-tuned state-of-the-art language models such as BERT Multilingual Cased and AfroLM to capture features in Yoruba proverbs. [1]

2 Yoruba Language

The Yoruba language belongs to the Niger-Congo language family and is spoken principally in southwestern Nigeria and parts of the Republic of Benin and Togo. It has about 25 million mother-tongue speakers. The language has 25 letters of the Latin alphabet including additional letters contain subdots, such as (e., gb, s., and o.). Yoruba is a tonal language, meaning that it has three distinctive tone levels-high, mid, and low-that are decisive in word distinction. Accurate pronunciation

[1] https://huggingface.co/datasets/LingoJr/OWEYOR

Figure 1: The deployed model tested with Yoruba proverbs and non-proverbs

depends greatly on the tonal marks and subdots. Proverbs have an important place in Yoruba culture, serving as crucial instruments of communication, wisdom, and social principles. In fact, their usefulness in conveying meaning and in the promotion of cultural identity in Yoruba society can hardly be underestimated. It must be mentioned that Proverbs serve as a vehicle for the transmission of wisdom, the resolution of conflicts, and the maintenance of identity among Yorubas. Successful classification of proverbs requires familiarity with patterns of language, difference in tone, and situational use.

3 Related Work

Numerous scholars have worked on various aspects of Yoruba proverbs, exploring their meaning, cultural significance, and application. Examples of these works are (Olubode Sawe, 2009), where the paper investigates how listeners derive meaning when they hear a proverb in conversation and the cognitive and cultural strategies used by listeners to arrive at the metaphorical or contextual meaning of proverbs. The paper also takes into account the linguistics cues embedded in proverbs that aid interpretation. (Alabi, 2015) discussed the role of proverbs as a stylistic and rhetorical device in communicating themes, expressing cultural identity, and addressing sociocultural and linguistic factors that influence the use of proverbs in literature. (Ademowo and Balogun, 2014) examines how the Yoruba proverb is used as a tool to promote peace and harmonious coexistence, managing interpersonal and communal conflicts, serving as a vehicle for teaching societal values and conflict resolution strategies. The study is centered on 24 randomly selected Yoruba proverbs that address themes of warning, cooperation, and diversity, and their implications for conflict management. Most of these works mentioned above are not use for any NLP tasks.

While previous studies have greatly explored various aspects of Yoruba proverbs, their application in NLP tasks still remains limited. However, recent studies in other languages show the potential of integrating proverbs in NLP task, part of which are (Liu et al., 2024), where the paper examines how multilingual large language model recognizes proverbs and sayings across different languages and cultures. The study introduced a dataset called MAPS (MulticulturalAI Proverb and Saying) designed to evaluate mLLMs. English, German, Russian, Bengali, Mandarin Chinese, and Indonesian are the languages covered. (Anita and Subalalitha, 2011) also worked on a system to automatically generate meanings for Tamil proverbs using a sentence scoring approach. This research is particularly significant as many Tamil proverbs are either misunderstood or lack accessible explanations, especially for the younger Tamil-speaking generation. (Baptista and Reis, 2022) investigate the use of Natural Language Processing (NLP) and Machine Learning (ML) techniques for thematic classification of Portuguese proverbs, which is the closet study to ours but what make it different is the language in discourse. [2]

4 ÒWE-YOR Dataset Collection

Textual Dataset: ÒWE-Yoruba was developed to try and bridge the gap in the lack of culturally significant datasets for low-resource languages. This paper deals with Yoruba language. In this paper, a classification task is proposed:

[2] https://huggingface.co/spaces/Joycenaomi81/Proverbs_

the Yoruba proverb or non-proverb classification dataset. These proverbs were sourced from an online PDF document with over 7,963 Yoruba proverbs, alongside their translations in English and meanings. In scraping, some of the sentences lost their diacritics and orthographic errors were seen in some of the sentences. The error mentioned was carefully corrected through manual editing to keep the linguistic integrity of the dataset intact.

The non-proverb portion of the dataset is made up of 7,962 sentences developed by native Yoruba speakers. These are on varied topics and include both everyday conversations and politically inclined statements to make sure that the representation of the non-proverbial text is wide. This step focused on cultural accuracy and linguistic diversity to complement the proverbs. The linguist annotating the data is a native speaker of Yoruba with deep knowledge of the culture and tradition. He has been involved in correct classification of data as either a proverb or non-proverb. Extensive cleaning, verification, and validation were done to guarantee the quality and reliability of this data.

Altogether, OWE-Yoruba contains 15,925 sentences for the textual data, with exactly 7,963 proverbs and 7,962 non-proverbs. This balanced dataset is a very useful resource in the development and evaluation of natural language processing models, particularly those tasks that require the presence of linguistic and cultural nuances in Yoruba.

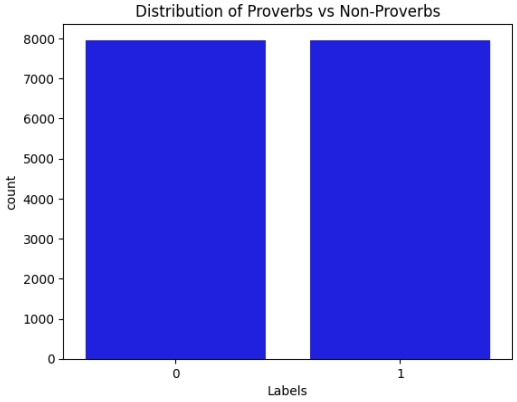

Figure 2: Visualization of the collected dataset

5 Experiment and Result

In this study, we trained a Naive Bayes model and fine-tuned AfroLM (Dossou et al., 2022) as well as a multilingual BERT-based model (Devlin et al., 2019) for the detection of Yoruba proverbs. Since AfroLM and BERT are designed to handle low-resource languages such as Yoruba, they are suitable candidates for this classification task.

The Multinomial Naive Bayes classifier performed at 85% on the test set. Its confusion matrix, therefore, showed that there were 1,449 true positives, 1,286 true negatives, 171 false positives, and 279 false negatives. While somewhat powerful, the Multinomial Naive Bayes classifier still couldn't make a reasonably accurate discrimination between proverbs and non-proverbs. On the other hand, fine-tuning AfroLM signifi- cantly improved this to 95% accuracy. From the confusion matrix for AfroLM, there were 1,511 true positives, 1,522 true negatives, 62 false pos- itives, and 92 false negatives. This improve- ment underlines the advantages of fine-tuning pre- trained models on specific tasks. The best result among the BERT multilingual base models achieved an accuracy of 96% on the test set. Its confusion matrix shows 1,565 true positives, 1,515 true negatives, 69 false positives, and 38 false negatives. It goes to show that the model does very well in the complexity of Yoruba proverb classification. In other words, the results reflect that, though Naive Bayes is a simple machine learning baseline, the models like AfroLM and BERT, which have been pre-trained and fine-tuned on this particular task, outperform all the other approaches by a great margin. This shows how well this model captures nuanced linguistic and cultural features. The very high accuracy of these models indicates a possibility for advanced tool development in low-resource languages.

Model	TP	TN	FP	FN
Naive Bayes	1,449	1,286	171	279
AfroLM	1,511	1,522	62	92
BERT	1,565	1,515	69	38

Table 1: Confusion matrix values for Naive Bayes, AfroLM, and BERT.

6 Conclusion and Future work

This work focuses on the creation of culturally relevant datasets and uses them for classification tasks that leverage the application of state-of-the-art machine learning methods for low-resource languages such as Yoruba. The results show

Model	Accuracy
Multinomial Naive Bayes	85%
AfroLM (fine-tuned)	95%
BERT multilingual cased (fine-tuned)	96%

Table 2: Accuracy of the machine learning baseline and fine-tuned models for Yoruba proverb detection.

that both traditional methods, including the Naive Bayes classifier, and fine-tuned pre-trained language models, including AfroLM and BERT multilingual base, are effective in classifying Yoruba proverbs. The machine learning model that was trained proved much more accurate, underlining their adaptability and potential in tasks requiring deep linguistic and cultural understanding.

Building on this work, We plan to extend the proverb dataset to other African languages and also curate speech dataset to train existing speech model capable of interpreting linguistics and cultural nuances of Yoruba proverbs. Speech datasets are extremely important for a variety of applications in speech recognition and culturally sensitive conversational AI. This work can also be easily scaled to other African languages through the creation of culturally inclined datasets for a variety of NLP and AI tasks, encouraging inclusivity and linguistic representation in technology.

7 Acknowledgment

We would also like to extend our sincerest and most profound thanks to Aremu Anuoluwapo for the priceless mentorship and support throughout the process. Secondly, the late Professor Oyekan Owomoyela for such a great feat in ensuring Yoruba proverbs live on through digital means. Finally, our sincere thanks go to Professor Ayo Yusuff, a renowned linguist and Research Professor at the Institute of African and Diaspora Studies, University of Lagos, for his inspiring words of encouragement to bring low-resource languages into the forefront of AI advancements.

References

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022a. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States, July. Association for Computational Linguistics.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O Alabi, Shamsuddeen H Muhammad, Peter Nabende, et al. 2022b. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. *arXiv preprint arXiv:2210.12391*.

Adeyemi Johnson Ademowo and Noah Balogun. 2014. Proverbs and conflict management in africa: A study of selected yoruba proverbs and proverbial expressions. *International Journal of Literature, Language and Linguistics*, 1(1):1–7.

Taofiq Adedayo Alabi. 2015. The poetics of yoruba proverbs in nigerian literature in english.

R Anita and CN Subalalitha. 2011. Automatic generation of description for tamil proverbs.

Jorge Baptista and Sónia Reis. 2022. Automatic classification of portuguese proverbs. In *11th Symposium on Languages, Applications and Technologies (SLATE 2022)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Bonaventure F. P. 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.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure FP Dossou, Akintunde Oladipo, Doreen Nixdorf, et al. 2023. Masakhanews: News topic classification for african languages. *arXiv e-prints*, pages arXiv–2304.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Rubungo Andre Niyongabo, Qu Hong, Julia Kreutzer, and Li Huang. 2020. KINNEWS and KIRNEWS: Benchmarking cross-lingual text classification for Kinyarwanda and Kirundi. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5507–5521, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

FO Olubode Sawe. 2009. Interpreting yoruba proverbs: Some hearer strategies. *California Linguistic Notes*, 34(2):1–17.

Daniele Pighin, Carlo Strapparava, et al. 2019. A proverb is worth a thousand words: Learning to associate images with proverbs. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 41.

Amanda P Rassi, Jorge Baptista, and Oto Vale. 2014. Automatic detection of proverbs and their variants. In *3rd Symposium on Languages, Applications and Technologies*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.