# YOSM: A NEW YORÙBÁ SENTIMENT CORPUS FOR MOVIE REVIEWS

**Anonymous authors**
Paper under double-blind review

**Sentiment Analysis** is a popular text classification task it involves applying natural language processing approaches to distinguish sentiment in texts with machine language models. The results of this task can be used by business owners and product developers to understand their consumer's perceptions about their products. Asides from customer feedback and product/service analysis, this task can be useful for social media monitoring (Martin et al., 2021). Sentiment analysis is well known for classifying and detecting the positive and negative sentiments on movie reviews. Movie reviews enable movie producers to monitor the performances of their movies (Abhishek et al., 2020) and enhance the decision of movie viewers to know whether it is risky to invest their precious hours in a bad movie or a movie that is good enough worth investing time to watch (Lakshmi Devi et al., 2020). However, the task has been under-explored on African languages compared to their western counterparts, "*high resourced languages*", that are privileged to have received enormous attention due to the enormous amount of available textual data. African languages fall under the category of the low resourced languages are on the disadvantaged end because of the limited availability of data that gives them a poor representation (Nasim & Ghani, 2020). Recently, sentiment analysis has received attention on African languages in the Twitter domain for Nigerian (Muhammad et al., 2022) and Amharic (Yimam et al., 2020) languages. However, there is no available corpus in the movie domain. For this reason, we decided to fill and tackle the problem of unavailability of Yorùbá data for movie sentiment analysis, we created the first Yorùbá sentiment corpus for Nollywood movie reviews. Also, we develop sentiment classification models using the state-of-the-art pre-trained language models like mBERT (Devlin et al., 2019) and AfriBERTa (Ogueji et al., 2021).

**Yorùbá Language** is the third most spoken indigenous African language (Eberhard et al., 2020) with over 50 million speakers. Speakers of the Yorùbá language can be found in the South-Western region of Nigeria and beyond the globe. Yorùbá is a tonal language and its alphabet comprises 25 letters. The language uses diacritics on and under dots, which are necessary for the understanding of the Yorùbá texts. Despite its large number of speakers, Yorùbá falls under the category of the low resourced languages and there are few NLP datasets that have been developed for the language (Adelani et al., 2021). Furthermore, there is no record of sentiment analysis research done on Nigerian movies (i.e. *Nollywood*) or even Yorùbá movie reviews.

**Nollywood** is the home for Nigerian movies that depict the Nigerian people and reflect the diversities across Nigerian cultures. Records show that Nollywood is the second-largest movie and film industry in the world. A Masterclass staff, Foster in 2022 [1], claims that four to five movies are released daily by Nigerian movie producers for an estimated audience of fifteen million Nigerians and five million in other African countries. Despite its capacity, Nollywood movie reviews are scarce.

**Data**: Sourcing for reviews of Nigerian movies written by Nigerians for Nigerians was cumbersome because Nigerians hardly write reviews. Unlike Hollywood movies that are heavily reviewed with hundreds of thousands of reviews all over the internet, there are fewer reviews about Nigerian movies. Furthermore, there is no online platform that purely has movie reviews originally written in Yorùbá. Most of the reviews are written in English. We collected 1500 reviews with a balanced set of positive and negative reviews. These reviews were sourced from three popular online movie database platforms [2] - IMDB, Rotten Tomatoes and Letterboxd. We also collected some reviews from two Nigerian indigenous movie reviews websites [3] - Cinemapointer and Nollyrated. Our annotation focused on the classification of the reviews based on the ratings that the movie reviewer gave the movie. We used a rating scale to classify the positive or negative reviews and defined ratings between 0-4 under the negative (NEG) category while 7-10 were positive (POS). After collecting the

[1] https://www.masterclass.com/articles/nollywood-new-nigerian-cinema-explained
[2] www.imdb.com , www.rottentomatoes.com, and https://letterboxd.com/
[3] www.cinemapointer.com, and https://nollyrated.com/

| Sentiment | No. Reviews | Ave. Length (No. words) | Data source | | | | |
|---|---|---|---|---|---|---|---|
| | | | IMDB | Rotten Tomatoes | LetterBoxd | Cinemapoint | Nollyrated |
| positive | 750 | 73 | 402 | 105 | 81 | 101 | 61 |
| negative | 750 | 63 | 278 | 133 | 101 | 193 | 46 |

Table 1: Data source, number of movie reviews per source, and average length of reviews

| Model | F1-score |
|---|---|
| mBERT | $83.2_{\pm 1.8}$ |
| mBERT+LAFT | $86.2_{\pm 1.3}$ |
| AfriBERTa | $\mathbf{87.2_{\pm 0.6}}$ |

(a) Benchmark results

| Model | Transfer learning setting | | | |
|---|---|---|---|---|
| | imdb (en) | en | yo:MT | en+yo:MT |
| mBERT | 61.4 | 61.9 | 71.5 | 74.0 |
| mBERT+LAFT | **69.4** | **71.8** | 76.5 | **77.9** |
| AfriBERTa | 64.3 | 69.8 | **77.1** | **77.9** |

(b) Transfer learning (F1-score)

Table 2: Benchmark and transfer learning results (F1-score). All results are average over 5 runs except transfer from "imdb".

data, two native speakers of Yorùbá that work as professional translators were recruited to manually annotate and translate the movie reviews from English to Yorùbá. We also automatically translate the English reviews to Yorùbá using Google Translate machine translation tool, this can be useful for scenarios where there is absence of training data in Yorùbá language. Table 1 shows the information about the data sources of the curated Yorùbá movie reviews, which we named YOSM. We split YOSM into 800 reviews as training set, 200 reviews as development set and 500 reviews as test set.

**Baseline Models** We *fine-tune* two pre-trained language models (PLMs) that have been pre-trained on Yorùbá language: mBERT (Devlin et al., 2019) and AfriBERTa (Ogueji et al., 2021). AfriBERTa has been exclusively pre-trained on 11 African languages while mBERT was pre-trained on 104 languages. As an additional baseline model, we make use of a PLMs that has been adapted to Yorùbá language using language adaptive fine-tuning (LAFT) – an approach to fine-tune PLM on monolingual texts in a new language using the same masked language model objective as BERT. It has been shown to improve performance on named entity recognition task on Yorùbá (Alabi et al., 2020; Adelani et al., 2021) and better zero-shot cross-lingual transfer (Pfeiffer et al., 2020).

**Transfer learning setting** We examine four transfer learning experiments, (1) **imdb (en)**: cross-lingual transfer from a large Hollywood movie review dataset (i.e IMDB) with 25,000 samples and zero-shot evaluation on YOSM test set. (2) **en**: cross-lingual transfer from the English Nolloywood movie review – the size is limited to the 800 samples in the untranslated reviews in our dataset. (3) **yo:MT**: trained on machine translation of 800 English Nolloywood reviews to Yorùbá language. (4) **en+yo:MT** Combined the English Nolloywood reviews and machine translated reviews.

**Results** Table 2 shows the **baseline results** on PLMs, we obtained very impressive results ($> 83$ F1) by training on our small training set (i.e 800 reviews). AfriBERTa and mBERT+LAFT gave better results (more than 86 F1) compared to mBERT (83.2) since they have been trained exclusively on African languages or adapted using LAFT. For the **transfer learning results**, we obtained a very good cross-lingual transfer of over (61 F1) on all settings. We find the transfer of **en** to perform better than **imdb(en)**, an improvement on of $2.4 - 5.5$ F1 using mBERT+LAFT or AfriBERTa since **en** captures better the Nollywood domain better than **imdb(en)** that is based on Holloywood reviews. The best transfer approach in the absence of humanly written Yorùbá reviews is to train on machine translated reviews (**yo:MT**) and/or combine with English Nolloywood reviews (**en+yo:MT**), with performance reaching 77.9 F1. There is small benefit of combining english and Yorùbá nollywood reviews ($0.8 - 2.5$ F1) to further improve performance.

**Conclusion** In this paper, we presented the first Yorùbá sentiment corpus for Nollywood movie reviews - YOSM that was manually translated from English Nollywood reviews. We perform experiments on this dataset by using the state-of-the-art pre-trained language models and transfer learning approaches which gave us impressive results.

## REFERENCES

Kumar Abhishek, Mayank Mehiral, and M. S. Sathvik Murthy. Sentimental analysis for movie reviews. *International Journal of Advanced Research in Computer Science*, 11(0):17–22, 2020. ISSN 0976-5697. doi: 10.26483/ijarcs.v11i0.6536. URL http://www.ijarcs.info/index.php/Ijarcs/article/view/6536.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9: 1116–1131, 2021. doi: 10.1162/tacl_a_00416. URL https://aclanthology.org/2021.tacl-1.66.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina España-Bonet. Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 2754–2762, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.335.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

B. Lakshmi Devi, V. Varaswathi Bai, Somula Ramasubbareddy, and K. Govinda. Sentiment analysis on movie reviews. In P. Venkata Krishna and Mohammad S. Obaidat (eds.), *Emerging Research in Data Engineering Systems and Computer Communications*, pp. 321–328, Singapore, 2020. Springer Singapore. ISBN 978-981-15-0135-7.

Gati Lother Martin, Medard Edmund Mswahili, and Young-Seob Jeong. Sentiment classification in swahili language using multilingual bert. *ArXiv*, abs/2104.09006, 2021.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Abdullahi Salahudeen, Aremu Anuoluwapo, Alípio Jeorge, and Pavel Brazdil. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. *ArXiv*, abs/2201.08277, 2022.

Zarmeen Nasim and Sayeed Ghani. Sentiment analysis on urdu tweets using markov chains. *SN Comput. Sci.*, 1:269, 2020.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 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*, pp. 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.11. URL https://aclanthology.org/2021.mrl-1.11.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on*

*Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7654–7673, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.617. URL https://aclanthology.org/2020.emnlp-main.617.

Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew Ayele, and Chris Biemann. Exploring Amharic sentiment analysis from social media texts: Building annotation tools and classification models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 1048–1060, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.91. URL https://aclanthology.org/2020.coling-main.91.