ENHANCING TRANSFORMER MODELS FOR IGBO LANGUAGE PROCESSING:A CRITICAL COMPARATIVE STUDY

Anthony Soronnadi, Olubayo Adekanmbi & Chinazo Anebelundu
Data Scientists Network
{anthony, olubayo, chinazo}@datasciencenigeria.ai

David Ifeoluwa Adelani
University College London
d.adelani@ucl.ac.uk

ABSTRACT

This paper reports on an ongoing investigation aimed at reviewing and optimizing Transformer models for processing the African language Igbo, which has limited resources. Creating an effective language model is essential for enhancing NLP applications in this setting, given the specific challenges posed by Igbo's rich morphological structure, tonal system, and limited availability of digital resources. In order to investigate the adaptation and optimization of Transformer models and to improve the models for Igbo language processing, this work takes a critical comparison approach. First efforts have focused on developing a RoBERTa model pre-trained on clean Igbo text corpus, and evaluating its performance on downstream tasks such as named entity recognition, text classification, and sentiment analysis. In our evaluations across the above-mentioned NLP tasks, IgboBERTa demonstrates competitive or superior performance relative to larger models such as XLM-R-large, XLM-R-base, AfriBERTa, and AfroXLMR-base, particularly when considering its efficiency due to its smaller size of only 83.4M parameters. This efficiency makes IgboBERTa particularly appealing for resource-constrained environments common in African NLP applications.

Introduction Significant progress has been made in natural language processing (NLP) in recent years, mostly due to the development and implementation of Transformer models, which have transformed how machines comprehend and produce human languages (Vaswani et al., 2017). But not everyone has benefited equally from these technological advances, with languages with a wealth of digital resources seeing the greatest advancements. Speaking to millions of people in Nigeria and beyond, the Igbo language stands in sharp contrast as one with fewer resources and one that has not completely benefited from the most recent advances in NLP (Warstadt et al., 2020). Creating efficient language models presents special difficulties because Igbo is a complex language and there are few digital resources available. This difference highlights the need for NLP technology to be customized to meet the unique requirements and traits of languages with fewer resources.

Similar initiatives have been undertaken to enhance NLP resources for Igbo, a language with limited digital assets. For instance, a study by Chukwuneke et al. (2022) presents the development of an Igbo-named entity recognition (NER) dataset and the experimentation with IgboNER models. It notably does not provide publicly accessible models for community evaluation Chukwuneke et al. (2022). Our work complements these efforts by not only detailing the development of an Igbo language model but also ensuring the availability of our models and datasets to the public on Hugging Face[1] and GitHub[2]. This approach emphasizes the importance of transparency and reproducibility in research, allowing for broader contributions to Igbo NLP and the advancement of technology for under-resourced languages.

[1] https://huggingface.co/DSNResearch/IgboBERTa
[2] https://github.com/DataScienceNigeria/IgboBERTa

Dataset	Task
MasakhaneNER2.0	Named Entity Recognition
Masakhana News Classification	Text Classification
AfriSenti	Sentiment Analysis

Table 1: Overview of Datasets and Their Tasks

Our research aims to evaluate and enhance Transformer models for Igbo language processing in order to close this gap. The use of Transformer models, like RoBERTa (Liu et al., 2019), is encouraged by their cutting-edge capabilities in a variety of natural language processing tasks for well-resourced languages. This study uses a critical comparison methodology to investigate optimization strategies for these models as well as evaluate how well they adapt to Igbo culture. In order to empower the Igbo-speaking people digitally, we want to bridge the technological gap by creating a strong Igbo language model that can enable a broad range of NLP applications.

We have focused our early efforts on assessing the adaptability and performance of the RoBERTa model on important NLP tasks, such as named entity recognition, text categorization, and Igbo-specific translation. Preliminary results show notable improvements in model performance. This study describes the approach used to make these changes, the difficulties encountered in obtaining and preparing Igbo language datasets, and the implications of our results for the use of NLP in less resource-rich linguistic situations in the future.

Pre-trained Dataset Our data collection process combined automated web scraping with random manual review and curation. To ensure the dataset's quality and diversity, heuristic-based quality filtering and rigorous de-duplication Zhao et al. (2023) were applied, enabling us to only include high-quality texts. This rigorous process underpins the enhanced learning material for the model, ensuring a foundation that mirrors the richness of the Igbo language.

The adaptation of a custom Byte-Pair Encoding (BPE) tokenizer, specifically developed for our Igbo corpus with a vocabulary size of 52000 was pivotal in addressing the language's unique agglutinative structure and morphological depth. This tokenizer plays a crucial role in minimizing out-of-vocabulary (OOV) tokens, ensuring a better-nuanced representation of features of the Igbo language for NLP tasks.

Our pre-training corpus, totaling 252 million tokens, was meticulously compiled from diverse sources, including web-scraped BBC Igbo articles, Jehovah's Witness publications, educational textbooks, and the NLLB machine translation corpus (NLLB-Team et al., 2022). This varied compilation strategy ensures a comprehensive linguistic representation, covering a broad spectrum of topics and styles that range from formal to conversational Igbo.

Pre-training Architecture The RoBERTa model, pre-trained with a Masked Language Modeling objective and masking 15% of the dataset, was optimized for the Igbo language. It features an architecture with 6 Transformer layers, 12 attention heads per layer, a vocabulary size of 52,000 with maximum position embeddings of 514, and 83 million parameters. This configuration enables efficient processing and deep learning of Igbo's linguistic nuances, ensuring precise language understanding.

Downstream Tasks and Datasets For downstream NLP tasks, we utilized the MasakhanaNER 2.0 (Adelani et al., 2022) for named entity recognition, MasakhaNEWS (Adelani et al., 2023) for text classification, and AfriSenti (Muhammad et al., 2023) for sentiment Analysis, as shown in Table 1, ensuring a comprehensive evaluation across varied NLP applications.

Experimental Setup and Technique For our experimental setup, we utilized 2 NVIDIA A10 GPUs , and we pretrained the RoBERTa model with 84 million parameters on the Igbo language, employing PyTorch and the Transformers library. The training involved a custom Igbo corpus with tailored preprocessing and a dynamic learning rate (1e-4). The pre-trained IgboBERTa model can be accessed on Hugging Face at https://huggingface.co/DSNResearch/IgboBERTa. For more details on fine-tuning IgboBERTa, kindly visit the project's GitHub repository at https://github.com/DataScienceNigeria/IgboBERTa.

Model	Size	NER F1-score	News Topic F1-score	Sentiment F1-score	Average
AfriBERTa (Ogueji et al., 2021)	126M	87.3	87.3	78.6	84.4
XLM-R-base (Conneau et al., 2020)	270M	87.8	82.5	75.6	82.0
XLM-R-large (Conneau et al., 2020)	550M	87.2	84.2	76.5	82.6
AfroXLMR-base (Alabi et al., 2022)	270M	88.5	90.7	76.3	85.2
AfroXLMR-large (Alabi et al., 2022)	550M	**89.6**	**93.4**	**79.5**	**87.5**
IgboBERTa (ours)	83M	87.4	89.3	78.5	85.0

Table 2: Comparison of IgboBERTa results with previous masked language models. We obtained baseline results from the papers of MasakhaNER 2.0, MasakhaNEWS, and AfriSenti. All models except XLM-R has seen Igbo during pre-training. IgboBERTa results were generated with an average of 5 runs for each task with different seeds for each run

The IgboBERTa model has demonstrated exceptional performance on their respective tasks, with a particular highlight on their effectiveness when evaluated using the F1 score as a metric. For text classification, the IgboBERTa, with a model size of 83M parameters, achieved an F1 score of 89.3 on the MasakhaNEWS topic classification dataset. This performance surpasses several larger and notable models on the same MasakhaNEWS classification dataset, including XLM-R-large (550M) with an F1 score of 84.2, XLM-R-base (270M) with 82.5, AfriBERTa (127M) with 87.3. However, it is slightly outperformed by AfroXLMR-base (270M), which scored 90.7. IgboBERTa also outperformed, AfriBERTa-large, XLM-R-base, AfroXLMR-base, and XLMR Large on MasakhaNEWS data (see Table 2).

Similarly, the IgboBERTa model has excelled in the named entity recognition task, achieving an overall F1 score of 86.5 on the MasakhaNER 2.0 dataset, which is very close to the performance of models like AfroXLM-R-large (89.6 F1 score) considering the size of both models. Through our work, we reinforce that quality data is of great importance in training AI models. The comparison underscores the efficiency and effectiveness of IgboBERTa in understanding the Igbo language, showcasing its superiority over models with two times larger sizes.

ACKNOWLEDGMENTS

This work was supported in part by Oracle Cloud credits and related resources provided by Oracle.

REFERENCES

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. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. *arXiv preprint arXiv:2210.12391*, 2022.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure FP Dossou, Akintunde Oladipo, Doreen Nixdorf, et al. Masakhanews: News topic classification for african languages. *arXiv preprint arXiv:2304.09972*, 2023.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4336–4349, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.382.

Chiamaka Chukwuneke, Ignatius Ezeani, Paul Rayson, and Mahmoud El-Haj. Igbobert models: Building and training transformer models for the igbo language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 5114–5122, 2022.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Un-

supervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

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, et al. Afrisenti: A twitter sentiment analysis benchmark for african languages. *arXiv preprint arXiv:2302.08956*, 2023.

NLLB-Team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672, 2022. URL https://api.semanticscholar.org/CorpusID:250425961.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin (eds.), *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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Alex Warstadt, Yian Zhang, Haau-Sing Li, Haokun Liu, and Samuel R Bowman. Learning which features matter: Roberta acquires a preference for linguistic generalizations (eventually). *arXiv preprint arXiv:2010.05358*, 2020.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.