Research Statement

David Ifeoluwa Adelani *December 8, 2022*

Objectives and Vision

My primary research objective is to develop machine learning models for **under-resourced languages** by leveraging text, speech, and other multi-modal resources. The goal is to improve **access to information** and enhance **human-machine communication** for these under-resourced languages. There are over 7,000 languages in the world [1], and many of them are under-represented in natural language processing (NLP) research. Current research focuses on a few languages, mostly from Western Europe and some languages in Asia like Mandarin, Japanese, and Arabic. While there is some progress in developing multilingual pre-trained language models (PLMs) to support the top 100 languages with online presence, this represents less than 2% of the languages. My research goal is to extend language technology to these under-represented languages.

There are several reasons for the under-representation of many languages in NLP research, some are **societal** such as the (1) language diversity of NLP researchers, (2) lack of government support, (3) weak language policies of many countries, and (4) colonialism – that resulted into suppression of many native languages.[1] The other reasons for the under-representation of many languages are **data and compute** related such as (1) lack of large unlabelled and labelled data, (2) limited compute for developing and evaluating NLP models based on large PLMs, and (3) limited coverage of under-resourced languages in multilingual representation models — required to build NLP models for several NLP tasks.

The limited coverage of under-resourced languages in PLMs is critical, because most NLP models make use of **transfer learning** to adapt to a new task, domain or language. In the widely use "pre-train and fine-tune" paradigm, performance is often limited for unseen and distant languages, and especially those that have less vocabulary overlap or a non-supported script. Creating a multilingual model to cover many languages is challenging due to the lack of unlabelled texts for under-resourced languages and the "curse of multilinguality" [2] — which makes it computationally infeasible to learn a good representation of many languages in a single model. Scaling PLMs to a larger size has been cited as one potential solution, but it is difficult for researchers from under-represented communities to benefit from large PLMs due to limited compute [3].

My research focuses on three directions: (1) **Scaling multilingual evaluations** of NLP models to under-resourced languages. This involves the collection of labelled data from typologically-diverse languages using a participatory research approach [4, 5] — where native speakers are involved in data collection and development of language technologies for their language. (2) Developing NLP models to improve **access to information** in under-resourced languages, this will cover knowledge-intensive tasks like question answering and information retrieval. (3) Developing NLP models to enhance **human-human or human-machine communications** in their native languages, which covers NLP tasks like machine translation and speech processing.

I hope that by investing in this area of research, we can achieve a more equitable application of language technologies to several languages of the world.

1 Current Achievements

Most of my past and current projects can be categorized in two parts: (1) Ethical considerations in developing NLP models where I focus on topics of privacy and prevention of misuse of large language models [6, 7, 8, 9], and (2) NLP for under-resourced African languages. This section outlines my past and current research achievements on the latter [5, 10, 11, 12, 13].

1.1 Development of labelled data for under-resourced African languages

A major line of my research during my PhD has been the development of labelled datasets for under-resourced languages with a focus on African languages. Before 2020, many African languages did not have publicly available labelled data and are barely represented in technology. Through collaboration with Masakhane – a

[1] Effects of *colonialism* include the strategy of putting languages in a hierarchy of prestige leading to suppression of many languages (`https://www.goethe.de/prj/zei/en/pos/22902448.html`), maintenance of colonial linguistic hierarchies by post-colonial successors, and native speakers' perception that their language is inferior to the dominant colonial language

grassroots movement whose mission is to strengthen NLP research in African languages, I led the creation of large-scale dataset creation for **21 languages** in three important NLP tasks. (1) **MasakhaNER** - a named entity recognition dataset for 10 African languages [5], which was later extended to 21 languages this year [10]. (2) **MAFAND-MT** - a machine translation (MT) dataset in the news domain for 21 African languages [11], and (3) **MasakhaPOS** - a part-of-speech dataset for 20 African languages.[2]

MasakhaNER was particularly special because it was done entirely by volunteers from the Masakhane community, and resulted in top journal/conference papers at TACL and EMNLP, as well as a *best paper award* at the AfricaNLP Workshop. I also collaborated in the development of other datasets on news topic classification [3, 12], sentiment classification datasets [14, 15], machine translation [16], and text-to-speech [17]. These datasets have encouraged the development of Africa-centric PLMs [12], and have been used to filter large web-crawled data used to train MT models for African languages [18]. In the recently completed WMT shared task for large-scale MT African languages [18], MAFAND-MT was the most used dataset, used by six out of eight teams because the dataset was of high quality. Also, the dataset collection has encouraged the **large collaboration of researchers** across the African continent – this has improved the representation of Africans publishing in top NLP conferences.

Language	Sentence
English	The Emir of Kano turbaned Zhang who has spent 18 years in Nigeria
Amharic	የካኖ ኢምር በናይጅርያ ፩፰ ዓመት ያሳለፈውን ዣንግ ዋና መሪ አደረጉት
Éwé	Kano ƒe Emir na wobla ta na Zhang si nɔ Nigeria ƒe 18 ɔŋŋ la
Hausa	Sarkin Kano yayi wa Zhang wanda yayi shekara 18 a Nigeria sarauta
Nigerian-Pidgin	Emir of Kano turban Zhang wey don spend 18 years for Nigeria
Swahili	Emir wa Kano alimvisha kilemba Zhang ambaye alikaa miaka 18 nchini Nigeria
Wolof	Emiiru Kanó dafa kaala kii di Zhang mii def Nigeria fukki at ak juróom ñett
isiZulu	U-Emir waseKano ubeke isigqoko kuZhang osechithe iminyaka engu-18 eNigeria.

Figure 1: **Example of the MasakhaNER dataset**

1.2 Adapting Pre-trained language Models to African Languages

Existing massively multilingual PLMs like XLM-R [2] are often trained on only about 100 languages with the most resources on a large web crawl. Despite, seeing 100 languages during pre-training, they often achieve a decent performance on several downstream tasks for both high-resourced and low-resourced languages. However, there is still a **large performance drop for languages unseen during pre-training**, especially African languages. Another challenge of using PLMs is their large parameter size and **hardware restrictions** for researchers in low-resource communities like Africa to run these models.

In this project [12], we address the two challenges of developing a highly-effective PLM with fewer parameters. We created a new Africa-centric PLM, **AfroXLMR** by performing language adaptation to multiple languages at once. We fine-tuned XLM-R on monolingual texts from the 17 most-resourced African languages and three high-resourced languages that are widely spoken on the continent (English, French, and Arabic) to encourage cross-lingual transfer learning. AfroXLMR is currently the **state-of-the-art PLM** for natural language understanding tasks. For **reducing the parameter size of PLMs**, we performed **vocabulary compression** [19] by first removing vocabulary tokens from the embedding layer that correspond to scripts not used by African languages before performing language adaptation, thus effectively reducing the model size by 50% with a slight drop in performance compared to the uncompressed model. Our paper was awarded a *best paper award* at COLING 2022 for this contribution.

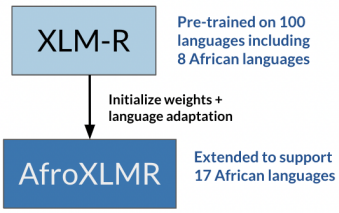

Figure 2: **AfroXLMR model**

1.3 Distant supervision and transfer learning for low-resource NLP

There are two popular definitions of low-resource NLP, one is **language-specific** based on how many digital resources are available, and the other is **task-specific** based on how many labelled data are available for a certain task (e.g NER) or domain (e.g NER for clinical texts). The latter can occur in any language.

Distant supervision can be used to create labelled data in a (semi-) automatic way. In the context of NER, it involves making use of automatic annotation rules by native speakers and matching lists of entities from an external knowledge source like gazetteers. To alleviate some of the negative effects of the errors in automatic annotation, noise-handling methods can be integrated with a few human-labelled examples [3]. The method has been shown to be effective for high-resourced languages like English.

[2] https://github.com/masakhane-io/masakhane-pos

We studied this approach on two under-resourced African languages (i.e. Hausa and Yorùbá) [13]. For automatic annotation of entities (i.e. personal name, organization, location, and date), we make use of an entity list from Wikidata and obtained *date* rules from native speakers based on date keywords like "ọjọ́" (day) and "oṣù" (month) in Yorùbá. Our evaluation shows that distant supervision is also effective for low-resource languages. However, we found that leveraging PLMs and transfer learning with a few examples (e.g. 10 or 100 sentences) from a high-resource language (e.g. English) to a low-resource language gives superior performance to distant supervision [3, 20]. The transfer learning approach has also been shown to be quite effective on other NLP tasks like machine translation, where a PLM fine-tuned on a few thousand parallel sentences in the target language or domain gives impressive performance [11, 21].

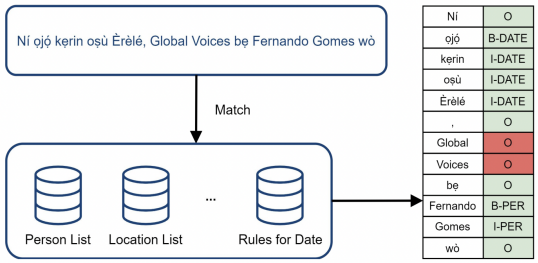

Figure 3: **Distant supervision for NER**

2 Future Directions

In order to achieve my research goals on under-resourced languages, I will focus on these three directions:

(1) Scaling up multilingual evaluation to under-resourced languages This involves extending the benchmarking of NLP models to more under-resourced languages. To overcome the lack of datasets, I plan to make use of non-traditional multilingual and multi-modal resources like large filtered web-crawled data (since they are typically noisy), unpaired speech e.g from YouTube, bilingual lexicons, and digitized printed books. A small number of labelled examples (e.g one hour of paired speech or 100 samples for Question answering (QA)) will be collected using a participatory research approach including standardized test sets. The focus will be on **creating a few samples across typologically-diverse languages** and developing efficient cross-lingual transfer learning methods for effective transfer to new tasks, domains, and languages.

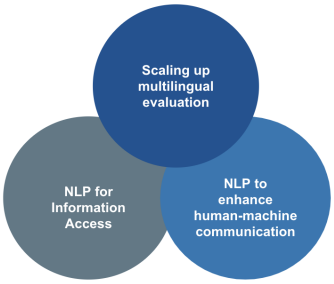

Figure 4: **Research Focus**

(2) NLP to enhance information access in under-resourced languages The focus will be knowledge-intensive tasks like open-domain QA and information retrieval. For open-domain QA, there is less focus on under-resourced languages because it requires a large, rich, and diverse external knowledge base like Wikipedia, but most languages have fewer articles in general on this platform. As summarized in [22], there are open challenges such as multi-step inference (retrieval and answer selection), and cross-lingual pattern matching (with or without a pivot language) to advance multilingual QA. Another important reason to focus on multilingual QA is that useful information about humanity is available in other cultures and languages (including under-resourced ones) that may not be present in English. There is an urgent need for QA systems that supports access to information in many languages, which can improve the digital literacy of communities worldwide.

(3) NLP to enhance human-human/human-machine communication This will be primarily focused on topics of machine translation (MT), text-to-speech, automatic speech recognition and speech translation. While there has been great progress in developing large-scale benchmark datasets for MT and speech recognition covering 100-200 languages like Flores-200 [23] and XTREME-S [24]. These large-scale datasets create an opportunity to advance research in both tasks. There are still other challenges of small and noisy training data for many languages. These datasets provide an opportunity to develop data-efficient methods [11]. In the short term, I would like to focus on a few open problems in MT for under-resourced languages such as developing appropriate **evaluation metrics** – since current embedding-based metrics with better correlation with human judgements only support languages covered by multilingual PLMs like XLM-R [25], building **parameter-efficient MT models** while retaining the performance of massively multilingual MT models, and developing effective methods for **adapting MT models to new domains**.

I believe my past and current research projects have prepared me for the research challenges ahead, I am looking forward to supervising graduate students on course projects and theses in these areas.

References

[1] M. P. Lewis, *Ethnologue: Languages of the world Sixteenth Edition.* SIL international, 2021. [Online]. Available: http://www.ethnologue.com/16/

[2] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *ACL*, 2020.

[3] M. A. Hedderich, D. Adelani, D. Zhu, J. Alabi, U. Markus, and D. Klakow, "Transfer learning and distant supervision for multilingual transformer models: A study on African languages," in *EMNLP*, Online, Nov. 2020, pp. 2580–2591.

[4] . ∀, W. Nekoto, V. Marivate, T. Matsila, T. Fasubaa, T. Fagbohungbe, S. O. Akinola, S. Muhammad, S. Kabongo Kabenamualu, S. Osei, F. Sackey, and et al., "Participatory research for low-resourced machine translation: A case study in African languages," in *Findings of EMNLP*, Online, 2020.

[5] D. I. Adelani, J. Abbott, G. Neubig, D. D'souza, J. Kreutzer, C. Lignos, C. Palen-Michel, H. Buzaaba, S. Rijhwani, S. Ruder, S. Mayhew, and et al, "MasakhaNER: Named entity recognition for african languages," *TACL*, 2021.

[6] D. Adelani, H. Mai, F. Fang, H. H. Nguyen, J. Yamagishi, and I. Echizen, "Generating sentiment-preserving fake online reviews using neural language models and their human- and machine-based detection," in *AINA*, 2020.

[7] A. A. Thomas, D. I. Adelani, A. Davody, A. Mogadala, and D. Klakow, "Investigating the impact of pre-trained word embeddings on memorization in neural networks," in *TSD*, 2020.

[8] D. I. Adelani, A. Davody, T. Kleinbauer, and D. Klakow, "Privacy guarantees for de-identifying text transformations," in *INTERSPEECH*, Oct. 2020.

[9] D. Adelani, M. Zhang, X. Shen, A. Davody, T. Kleinbauer, and D. Klakow, "Preventing author profiling through zero-shot multilingual back-translation," in *EMNLP*, Nov. 2021, pp. 8687–8695.

[10] D. I. Adelani, G. Neubig, S. Ruder, S. Rijhwani, M. Beukman, C. Palen-Michel, C. Lignos, J. O. Alabi, S. H. Muhammad, P. Nabende, ..., and D. Klakow, "MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition," *EMNLP*, 2022.

[11] D. Adelani, J. Alabi, A. Fan, J. Kreutzer, X. Shen, M. Reid, D. Ruiter, D. Klakow, P. Nabende, E. Chang, and at al., "A few thousand translations go a long way! leveraging pre-trained models for African news translation," in *NAACL-HLT*, Jul. 2022, pp. 3053–3070.

[12] J. O. Alabi, D. I. Adelani, M. Mosbach, and D. Klakow, "Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning," in *COLING*, Oct. 2022, pp. 4336–4349.

[13] D. I. Adelani, M. A. Hedderich, D. Zhu, E. van den Berg, and D. Klakow, "Distant supervision and noisy label learning for low resource named entity recognition: A study on hausa and yorùbá," *AfricaNLP*, 2020.

[14] S. H. Muhammad, D. I. Adelani, S. Ruder, I. S. Ahmad, I. Abdulmumin, B. S. Bello, M. Choudhury, C. C. Emezue, S. S. Abdullahi, A. Aremu, A. Jorge, and P. Brazdil, "NaijaSenti: A nigerian Twitter sentiment corpus for multilingual sentiment analysis," in *LREC*, Jun. 2022, pp. 590–602.

[15] I. Shode, D. I. Adelani, J. Peng, and A. Feldman, "NollySenti: Leveraging Transfer Learning and Machine Translation for Nigerian Movie Sentiment Classification," in *Under-submission*, 2023.

[16] D. Adelani, D. Ruiter, J. Alabi, D. Adebonojo, A. Ayeni, M. Adeyemi, A. Awokoya, and C. España-Bonet, "The effect of domain and diacritics in yoruba–english neural machine translation," in *MT Summit-2021*, 2021.

[17] J. Meyer, D. I. Adelani, E. Casanova, A. Oktem, D. W. J. Weber, S. K. Kabenamualu, E. Salesky, and et al., "BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus," in *INTERSPEECH*, 2022.

[18] D. I. Adelani, M. M. I. Alam, A. Anastasopoulos, A. Bhagia, M. R. Costa-jussà, F. Guzmán, H. Schwenk, and et al., "Findings of the WMT'22 Shared Task on Large-Scale Machine Translation Evaluation for African Languages ," in *Proceedings of the 7th Conference on Machine Translation at EMNLP*, 2022.

[19] A. Abdaoui, C. Pradel, and G. Sigel, "Load what you need: Smaller versions of mutililingual BERT," in *SustaiNLP at ACL*, Online, Nov. 2020, pp. 119–123.

[20] D. Zhu, M. A. Hedderich, F. Zhai, D. I. Adelani, and D. Klakow, "Task-adaptive pre-training for boosting learning with noisy labels: A study on text classification for african languages," *AfricaNLP*, 2022.

[21] E.-S. Lee, S. Thillainathan, S. Nayak, S. Ranathunga, D. Adelani, R. Su, and A. McCarthy, "Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation?" in *Findings of ACL*, May 2022.

[22] A. Asai, S. Longpre, J. Kasai, C.-H. Lee, R. Zhang, J. Hu, I. Yamada, J. H. Clark, and E. Choi, "MIA 2022 shared task: Evaluating cross-lingual open-retrieval question answering for 16 diverse languages," in *Multilingual Information Access @NAACL*, Jul. 2022, pp. 108–120.

[23] N. team, M. R. Costa-jussà, J. Cross, O. cCelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, and et.al., "No language left behind: Scaling human-centered machine translation," *ArXiv*, vol. abs/2207.04672, 2022.

[24] A. Bapna, C. E. Rivera, D. van Esch, J. Riesa, J. Clark, M. Johnson, M. S. Kale, M. Ma, O. Firat, S. Ritchie, S. Ruder, S. Khanuja, Y. Jia, and Y. Zhang, "Xtreme-s: Evaluating cross-lingual speech representations," in *Proc. Interspeech 2022*, 2022.

[25] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for MT evaluation," in *EMNLP*, Online, Nov. 2020, pp. 2685–2702.