FINAL TECHNICAL REPORT / RAPPORT TECHNIQUE FINAL

OUTREACH PROGRAMME TO STRENGTHEN THE AI4D NETWORK - FINAL TECHNICAL REPORT

Davor Orlic;

John Shawe-Taylor;Kathleen Siminyu;

IDRC Grant / Subvention du CRDI: 109187-002-Laying the foundations for artificial intelligence for development (AI4D) in Africa

Outreach Programme to strengthen the AI4D Network

Final Technical Report

IDRC Project Number-Component Number: 109187-002
By: Davor Orlic
Report Type: Final Technical Report
Period covered by the report: 1 January 2019 to 28 February 2021
Date: February 28 2021
Country/Region: United Kingdom
Full Name of Research Institution: Knowledge 4 All Foundation
Address of Research Institution: Betchworth House, 57-65 Station Rd, Redhill, Surrey, RH11DL
Members of Research Team: Davor Orlic, Kathleen Siminyu, John Shawe-Taylor
Contact Information of Research Team members: davor.orlic@ijs.si, kathleensiminyu@gmail.com, j.shawe-taylor@ucl.ac.uk

1. Executive Summary

The AI4D Outreach Programme provides support for the existing AI for Development (AI4D) Africa project designing a Network of Excellence in AI in sub-Saharan Africa to strengthen and develop community scientific and technological excellence in a range of AI-related issue areas. Specifically, the project:

1. Launched a Fellowship to develop datasets and strengthen capacities and innovation potential for Low Resource African Languages[1], including 29 researchers, covering 9 African languages, spoken across 22 countries, reaching 300 million speakers;
2. Launched a series of 5 data competitions with the mission of obtaining the best possible results using machine learning methods to solve challenges across African languages
3. Spurred 11 innovation projects[2] within the IndabaX community that focus on ethical, inclusive, participatory and gender-responsive approaches/applications to address development challenges, with consideration for SDG targets;
4. Received a Wikimedia Foundation Research Award of the Year 2021[3] for its combined efforts to support Masakhane[4].

AI4D Outreach Africa is complementary to the AI4D Network n. 108914-001 and result in the strengthening the establishment of the AI4D network which delivered a series of reports on African AI, an initial portfolio of language datasets and innovation projects, including recommendations for capacity building for ethical and locally relevant AI research around the African continent.

2. Research problem

The projects main approach and methodology have changed significantly from the initially proposed ones. The basic rationale and specific objectives of the project were to launch the AI4D Africa Network of Excellence conference, and a travel support programme for events organized within and outside the AI4D Network of excellence. These were not met, due to travel restrictions throughout the COVID-19 pandemic. Instead, funding was used to diversify the research portfolio in order to maximize on the already successful assets and best practices done initiated in grant n. 108914-001 and therefore allocated to a set of micro-projects and building language technologies for African languages. However, two other objectives remained the same, to extend the co-funding outside the AI4D Africa Network of Excellence and to launch the call for AI challenges to support an ecosystem of baseline datasets. The main objectives of supporting the burgeoning AI4D community in Africa and contributing towards building the AI4D field in Africa were reached. The contribution to knowledge that this project represents from a scientific, developmental and policy perspective are described in details Section 4, the research problems and general reflections are described below.

Problem 1: Capacity and community inside and outside the AI4D Africa Network of Excellence

In general, the problem being identified is that fewer African AI researchers and engineers result in fewer opportunities to use AI to improve the lives of Africans. The basic problem was how to support the growing AI4D Network and to minimize the so-called AI divide in Africa, support the already identified best practices in the current capacity in AI on the continent and based on the existing findings from the AI4D Network project, support meaningful events across Africa, to

[1] Cracking the Language Barrier for a Multilingual Africa https://www.k4all.org/project/language-dataset-fellowship/
[2] IndabaX and Data projects https://www.k4all.org/project/?type=international-development
[3] Wikimedia Foundation Research Award of the Year 2021 https://research.wikimedia.org/awards.html
[4] Masakhane github repository https://github.com/masakhane-io

contribute in making further informed decisions within the broader AI4D funding programme. This rationale has not changed in the past year, quite the opposite, it has proven to be more relevant than previously envisioned. African colleagues faced negative controversies in attending relevant international AI conferences like 2020 NeurIPS in Canada[5], consecutively their work was not being present on the international AI scene, with the COVID-19 pandemic slowing the creation of the AI4D Network. Therefore, we engaged with African AI communities in capacity building efforts via micro-projects in the main AI bottom-up communities such as Deep Learning Indaba, Data Science Africa and Data Science Nigeria, with the help of local IndabaX chapters across 30 African states.

Problem 2: Preservation of African Languages

As the rest of the world advances to integrate digital language and speech technologies into a variety of sectors the gap between high resource languages and those with less data becomes apparent and a key contributing factor to the ever-widening divide. In the context of language learning, for example, there exist numerous, freely available resources, applications and material, for language learners of different levels to begin or continue their journey, as appropriate. This is the case for major western languages and hardly the same for African languages. From a language preservation perspective, it is imperative that this work be undertaken. Increasingly, we find that urban Africans who have been born and/or raised in the city may not speak their mother tongue. Is unfortunate that in many such households, particularly where the parents are keen to equip their children for a better life, a greater emphasis is put on learning English, French or Portuguese, depending on where one is located, as parents have the perception that proficiency in these western languages are a more useful skill for career and life advancement in the world today. To solve this problem, we created the AI4D Language Dataset Fellowship program supporting the development of African language datasets for several Natural Language Processing (NLP) tasks. This work contributes to a roadmap for better integration of African languages on digital platforms in aid of lowering the barrier for African participation in the digital economy.

3. Progress towards milestones

All project milestones as specified in the Grant Agreement for the entire reporting period have been achieved. However, we briefly list here each main project milestones:

Milestone	Achievement evidence
Milestone 1: Launch travel support programme for events organized within and outside the AI4D Network of excellence	This milestone was not reached due to the COVID-19 pandemic. However, it was redefined and part of the funding was transferred to covering online registrations for and NLP workshop at ICLR 2020.
Milestone 2: Launch an AI4D conference in AI	This milestone was not reached due to the COVID-19 pandemic. However, it was postponed and part of the funding was transferred to a set of 11 micro-projects.
Milestone 3: Launch the call for AI challenges to support an ecosystem of baseline datasets	The team has successfully launched the following challenges: 1. AI4D iCompass Social Media Sentiment Analysis for Tunisian Arabizi (20 November 2020—29 March 2021): 539 Data scientists enrolled, 213 Data scientists on the leaderboard, 3 630

[5] Canada refuses visas to over a dozen African AI researchers https://www.bbc.co.uk/news/world-us-canada-50426774

	Submissions, Accuracy score: 0.94 (link) 2. AI4D Malawi News Classification Challenge (22 January—10 May): 218 Data scientists enrolled, 69 Data scientists on the leaderboard, 686 Submissions, Accuracy score: 0.64 (link) 3. AI4D Takwimu Lab – Machine Translation Challenge (18 December 2020—26 April 2021): 134 Data scientists enrolled, 11 Data scientists on the leaderboard, 142 Submissions, BLEU score: 0.35 (link) 4. AI4D Yorùbá Machine Translation Challenge (4 December 2020—12 April 2021): 314 Data scientists enrolled, 33 Data scientists on the leaderboard, 285 Submissions, BLEU score: 0.43 (link) 5. AI4D Baamtu Datamation - Automatic Speech Recognition in WOLOF (12 February—24 May) (link)
Milestone 4: Website support	The AI4D public website (link) was used for storing all information on the project activities.
Milestone 5: Call for proposals for micro-projects	The call for proposals has resulted in 109 responses with applications ranging from the following fields: Healthcare, Language, Agriculture, Governance, Education, etc. We have selected 11 projects that fitted the selection criteria, delivered via a ranking system in Baobab installation at the Deep Learning Indaba website and assessed by independent reviewers for each field. The projects kicked-off on September 15th 2020 and finished on February 28th 2021.
Milestone 6: Launch of dataset creation	The team has successfully launched the data creation for the following languages: 1. Ewe language (link) and Fongbe language (link) parallel text dataset (link) for Neural Machine Translation 2. Yoruba language (link) Machine Translation dataset 3. Chichewa language (link) document classification datasets (link) 4. Wolof language (link) text-to-speech dataset for 5. Kiswahili language (link) document classification datasets 6. Tunisian Arabizi language (link) sentiment analysis dataset (link) 7. Swahili: News Classification Dataset (link) 8. Twi language (link) 9. Luganda language (link)

4. Synthesis of research results and development outcomes

This project's overall objective is to build the foundations for responsible artificial intelligence for development work that contributes to sustainable development. Specifically, to support the burgeoning AI4D community in Africa and to contribute to building the AI4D field in Africa. Specifically, the projects current findings and results in the period 1 March 2020 to 28 February 2021 are complementary and connected to another IDRC grant number 108914-001 and combined have achieved the following outreach via their interconnected activities:

Reach of Artificial Intelligence 4 Development programme across Africa

In total 42 countries reached

Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon, Chad, Comoros, Congo, Democratic Republic of Congo, Egypt, Eswatini, Ethiopia, Gambia, Ghana, Guinea, Ivory Coast, Kenya, Lesotho, Madagascar, Malawi, Marocco, Mauritania, Mauritius, Mozambique, Namibia, Nigeria, Rwanda, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, Somalia, South Africa, Sudan, Tanzania, Tunisia, Togo, Uganda, Zambia, Zimbabwe

Main achievements

Supported 6 workshops
Commissioned 4 reports
Run 3 COVID-19 data challenges
Run 5 African Language data challenges
Funded 21 mini-projects
Launched a Fellowship for Low Resource African Languages
Developed 10 African language datasets
Built a text-to-speech platform for African Languages
Created a registry of AI hot spots in Africa
Engaged with ~ 150 researchers, ~ 30 institutions
Researched policy across 39 countries

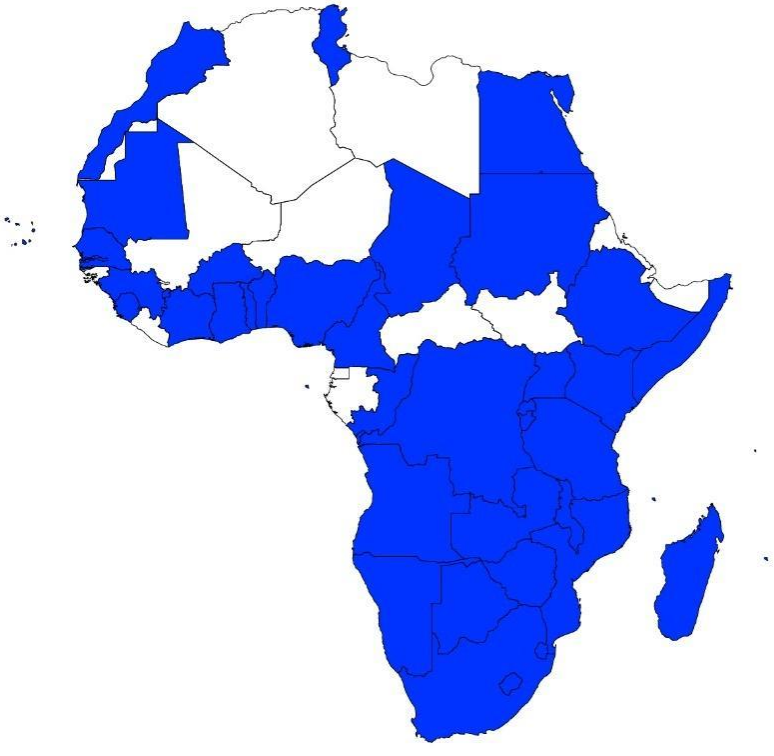

Figure 1: AI4D Network activities outreach across Africa in both

Milestone	Achievement evidence	Hard evidence
Micro-projects via innovation grants		
Micro-project	Characterizing Health Misinformation on Social Media	Access: • Project webpage • Presentation slides
Micro-project	AI system for MNC: Artificial intelligent system for predictors of early detection of maternal, neonatal and child health risks and their timely management	Access: • Project webpage • Project working website
Micro-project	AI for Coral Reef Conservation: Data Collection for Computer Vision in the Vamizi Island	Access: • Project webpage • and presentation slides

Micro-project	Development of Machine Learning Dataset for Poultry Diseases Diagnostics	Access: • Project webpage • Presentation slides • *Machine Learning Dataset for Poultry Diseases Diagnostics* dataset • GitHub data and code release • Journal abstract • Published blogpost • GitHub data and code
Micro-project	ChexNet Model Compression for Pneumonia Detection Using Low Powered Edge Devices	Access: • Project webpage • GitHub data and code release
Micro-project	Locally run Web-based App for Interpretable Breast Cancer Diagnosis from Histology Images	Access: • Project webpage • Presentation slides • Publication paper
Micro-project	Visual Question Answering in the Medical Domain	Access: • Project webpage • Presentation slides • GitHub data and code release
Micro-project	An African Short Story Language Corpus	Access: • Project webpage • Presentation slides • Project website • Additional website
Micro-project	Improving Online Learning Experience using Accent Transfer	Access: • Project webpage • Presentation slides • Publication paper • Published dataset • Data processing scripts on release
Micro-project	Computationally Accelerating Protein-Ligand Matching for neglected tropical disease	Access: • Project webpage • Presentation slides • ICLR 2021 paper • Indaba Grand Challenge: Curing Leishmaniasis by Deep Learning Indaba[6]
Micro-project	Keyword Spotting with African Languages	Access: • Data collection campaign call-to-action • Data collection tool link • Published dataset • GitHub data and code release • Deployed application platform • link to report • GitHub data and code

[6] Indaba Grand Challenge: Curing Leishmaniasis by Deep Learning Indaba https://zindi.africa/competitions/indaba-grand-challenge-curing-leishmaniasis

Milestone	Achievement evidence	Hard evidence
Awards		
Wikimedia Foundation Research Award of the Year	The Wikimedia Foundation Research team established the Wikimedia Foundation Research Award of the Year in 2021 to recognize recent research that has the potential to have significant impact on the Wikimedia projects or research in this space. The award was given to the "Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages and the Masakhane Community"	Access: • Paper here • and other papers here

Milestone	Achievement evidence	Hard evidence
Database creation		
Dataset creation	Ewe and Fongbe language	Access: • Ewe project webpage • Fongbe project webpage • Parallel text dataset for Neural Machine Translation
Dataset creation	Yoruba language	Access: • Yoruba project webpage • Machine Translation dataset
Dataset creation	Chichewa language	Access: • Chichewa project webpage • Document classification datasets
Dataset creation	Wolof language	Access: • Wolof project webpage • Text-to-speech dataset
Dataset creation	Kiswahili language	Access • Kiswahili project webpage • Document classification datasets
Dataset creation	Tunisian Arabizi language	Access: • Tunisian Arabizi webpage • Sentiment analysis dataset
Dataset creation	Twi language	Access: • Twi project webpage • Project dataset
Dataset creation	Luganda language	Access: • Luganda project webpage • Project dataset

Milestone	Achievement evidence	Hard evidence
Machine Learning Challenges		
GIZ AI4D Africa Language	This challenge hosted in partnership with GIZ and the FAIR Forward initiative and the	Access challenge here with 419 data scientists enrolled

Challenge - Round 2	Artificial Intelligence for Development Africa (AI4D-Africa) Network from 1 June 2020 to 3 August 2020	
Language Challenge - Round 2	AI4D iCompass Social Media Sentiment Analysis for Tunisian Arabizi (20 November 2020—29 March 2021)	539 Data scientists enrolled, 213 Data scientists on the leaderboard, 3 630 Submissions, Accuracy score: 0.94
Language Challenge - Round 2	AI4D Malawi News Classification Challenge (22 January—10 May)	218 Data scientists enrolled, 69 Data scientists on the leaderboard, 686 Submissions, Accuracy score: 0.64
Language Challenge - Round 2	AI4D Takwimu Lab – Machine Translation Challenge (18 December 2020—26 April 2021)	134 Data scientists enrolled, 11 Data scientists on the leaderboard, 142 Submissions, BLEU score: 0.35
Language Challenge - Round 2	AI4D Yorùbá Machine Translation Challenge (4 December 2020—12 April 2021)	314 Data scientists enrolled, 33 Data scientists on the leaderboard, 285 Submissions, BLEU score: 0.43
Language Challenge - Round 2	AI4D Baamtu Datamation - Automatic Speech Recognition in WOLOF (12 February—24 May)	Results due on May 25th
IndabaX micro-project challenge	Indaba Grand Challenge: Curing Leishmaniasis by Deep Learning Indaba (29 June 2020—1 June 2021)	312 data scientists enrolled, 24 on the leaderboard

Milestone	Achievement evidence	Hard evidence
Dissemination and Communication		
Publication	Presentation of common paper at NeurIPS 2020 workshop CiML 2020, ML Competitions at the Grassroots, AI4D - African Language Program in Ottawa, Canada	Access link
Publication	Presentation at WiNLP conference, co-located with ACL 2020	Access to ACL anthology link, and on Arxiv link

Result 1: Micro-projects and Innovation Applications

The IndabaX-AI4D Innovation Grants, which aimed to fund 6-month projects that support AI research communities and the work they do, especially during the COVID pandemic in 2020.

After a rigorous review process, 11 projects out of a total of 109 were selected. This grant programme has only been possible through deep partnership, and has been funded through a collaboration between the International Development Research Centre (IDRC), the Swedish International Development Cooperation Agency (SIDA), the Knowledge 4 All Foundation (K4A), and the International Research Centre on Artificial Intelligence under the auspices of UNESCO.

The IndabaX call for applications[7] was launched on June 12th 2020 and lasted for 6 weeks until July 6th. The projects selected for funding were notified on July 20th with an invitation to present their solutions at the AI4D workshop. The requirements for projects were:

1. that are conducted in Africa
2. that have a strong machine learning, artificial intelligence or data science component, in any discipline of science
3. that have sustainable development goals in mind
4. that could reach deliverable outcomes by the end of January 2021.

The selected projects were awarded from 4,000 - 8,000 USD which were disbursed in two funding rounds. Procedures for monitoring the project progresses, timelines and deliverables have been put in place. A highly effective mentorship structure was put in place as well. The following researchers were awarded:

1. Jeremiah Oluwaseye Fadugba, Independent researcher, AIMS Ghana Alumni
2. Mohamed Hassan Kane, Research Affiliate at MIT
3. Erwan Ciret Sola, Lúrio University, Mozambique
5. Volviane Saphir MFOGO, Independent / AIMS Cameroon Alumni
6. Tejumade Afonja, Saarland University
7. Rukayat Folashade Sadiq, Carnegie Mellon University
1. Ezinne Serah Nwankwo, The Nelson Mandela African Institution of Science and Technology, Duke University
8. Daouda Tandiang DJIBA, BICIS Group BNP Paribas, GalsenAI
9. Benson Muite, Independent researcher
10. Gladness G. Mwanga, Nottech Company Limited

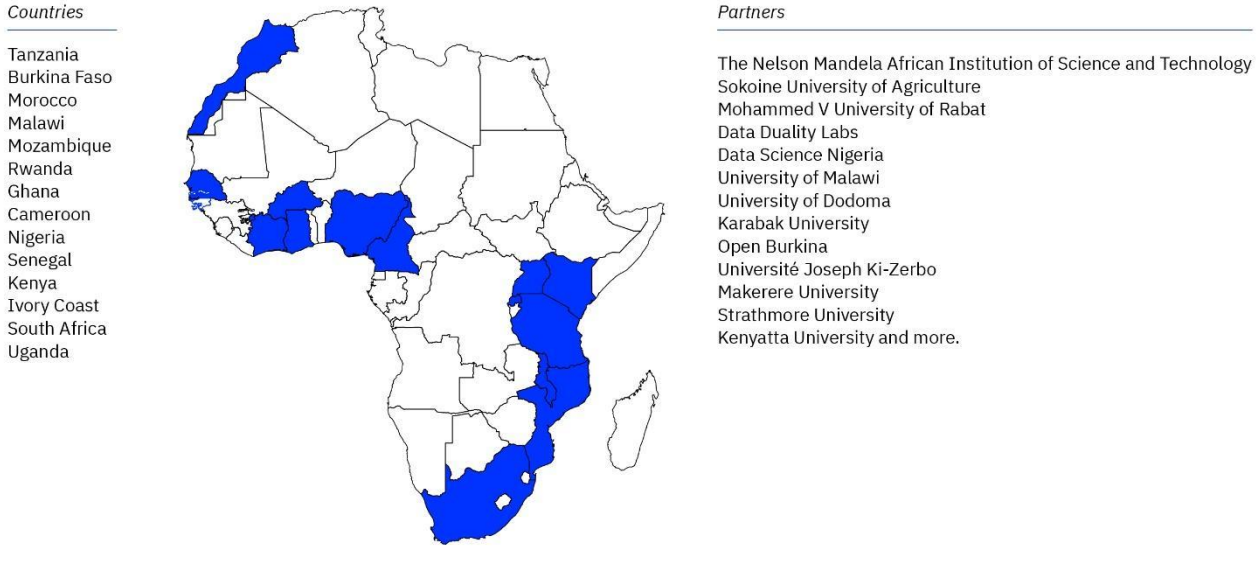

Figure 2: AI4D micro-projects across Africa

The 11 projects in AI within the Deep Learning Indaba community via IndabaX national chapters across Africa from healthcare, disease, language, education, agriculture, environmental conservation and ML solutions for the Indaba Grand Challenge on solving Leishmaniasis:

[7] https://deeplearningindaba.com/2020/ai4d-indabax-innovation-call-for-proposals/

1. **Characterizing Health Misinformation on Social Media** (webpage)
   - Quick summary: The objective of this project is to study the dynamics of the spread of factual and false information in online social networks in Nigeria during the pandemic.
   - Country: Nigeria
   - Team: Ofure Ebhomielen, Ezinne Nwankwo and Daniel Nkemelu

2. **AI for Coral Reef Conservation** (webpage)
   - Quick summary: The goal of this project is to develop a computer-vision based non-intrusive automatic data collection mechanism to collect images and give insights about coral reefs in the Vamizi Island and allow biologists to analyze data in real-time and infer on animals' life story, behaviour, population, and survivorship in Mozambican waters.
   - Country: Moçambique
   - Team: Erwan Sola and Luís Pina

3. **Development of Machine Learning Dataset for Poultry Diseases Diagnostics** (dataset and webpage)
   - Quick summary: The expected outcome of this work is to establish an annotated dataset for poultry diseases diagnostics for small to medium scale poultry farmers.
   - Country: Tanzania
   - Team: Hope Emmanuel Mbelwa, Ezinne Nwankwo, Dr. Dina Machuve, Dr. Neema Mduma and Dr. Evarest Maguo

4. **Visual Question Answering in the Medical Domain** (webpage)
   - Quick summary: This system takes as input a medical image and a clinically relevant question and outputs the answer based on the visual content.
   - Country: Cameroon
   - Team: Volviane Saphir MFOGO, Dr. Georgia Gkioxari, Dr. Xinlei Chen and Jeremiah Fadugba,

5. **Locally run Web-based App for Interpretable Breast Cancer Diagnosis from Histology Images** (webpage)
   - Quick summary: Wee will be building a Locally run web-based app for interpretable breast cancer diagnosis.
   - Country: Ghana
   - Team: Jeremiah Fadugba, Oluwayetunde Sanni and Moshood Olawale

6. **AI System for MNC (Maternal, Neonatal and Child Health** (webpage)
   - Quick summary: We will be building an AI system for predictors of early detection of maternal, neonatal and child health risks and their timely management.
   - Country: Tanzania
   - Team: Gladness G. Mwanga, Timothy Y. Wikedzi and Scott Businge

7. **Improving Online Learning Experience using Accent Transfer** (webpage)
   - Quick summary: This work will focus on making online educational content accessible through the reformulation of content in local accents.
   - Country: Nigeria
   - Team: Tejumade Afonja, Munachiso Nwadike, Olumide Okubadejo, Lawrence Francis, Clinton Mbataku, Femi Azeez and Wale Akinfaderin

8. **An African Short Story Language Corpus** (webpage)

- Quick summary: is intended to develop openly licensed free to use African language corpora.
- Country: Kenya
- Team: Prof. Audrey Mbogho, Dr. Lilian Wanzare, Dr. Benson Muite, Prof. Constantine Yuka and Mr. Juan Steyn

9. **Keyword Spotting with African Languages** (webpage)
   - Quick summary: The motivation of this work is to extend a speech commands dataset to include African languages, particularly focusing on 6 Senegalese languages: Wolof, Poular, Sérère, Mandingue, Diola, Soninké.
   - Country: Senegal
   - Team: Jean Michel Ahmath Sarr, Daouda Tandiang Djiba, Thierno Diop, Derguene Mbaye, Elias waly Ba, Ousseynou Mbaye and Dr Mamour Dramé

10. **ChexNet Model Compression for Pneumonia Detection Using Low Powered Edge Devices** (webpage)
    - Quick summary: The goal of this work is to build a model compression algorithm for ChexNet. The ChexNet network is chosen as the base model because it is the current state of the art technique in detecting Pneumonia on chest x-ray and as such, a reasonable choice.
    - Country: Rwanda
    - Team: Rukayat Sadiq, Brume Love, Jeremiah Fadugba, Olalekan Olapeju, Oluwafemi Azeez, Pelumi Oladokun and Tella Hambal

11. **Computationally Accelerating Protein-Ligand Matching for Neglected Tropical Diseases** (webpage)
    - Quick summary: We will be working on a solution for the Indaba Grand Challenge: Curing Leishmaniasis. The goal is to propose a new treatment, comprising a Leishmania protein (present in the proteome of one or more of the Leishmania species) and a small molecule (or set of small molecules).
    - Country: Ivory Coast and United States
    - Team: Kane Mohamed Hassan, Nkwate Ebenezer and Loic Kwate Dassi

Result 2: Database building

AI4D Africa Language Challenge - Round 1[8]

The *AI4D - Language Dataset Challenge* was conceptualized within the AI4D Network project (IDRC grant number 108914-001) as an effort to incentivize the creation, collation and uncovering of African Language datasets and was spear headed in this project. This 5-month process saw the submission of 35 datasets from a variety of African languages/dialects, among them Amharic, Ewe, Fongbe, Swahili, Twi, Wolof and Yoruba. In total 190 data scientists enrolled to solve the challenge.

GIZ, AI4D Africa Language Challenge and Fellowship programme - Round 2[9]

The second phase of the *AI4D - Language Dataset Challenge* has provided datasets for the 2nd Challenge. While the overall outcome of the 1st was overwhelmingly positive, one challenge encountered was the submission of small datasets given that evaluation was done on a monthly basis. In response and as a continuation of these efforts, this subsequent work will involve the selection of 5 teams, out of the 10 that emerged as winners during the initial challenge and inviting them to continue working on their datasets for a period of 5 extra months.

[8] AI4D Africa Language Challenge https://zindi.africa/competitions/ai4d-african-language-dataset-challenge
[9] GIZ AI4D Africa Language Challenge - Round 2https://zindi.africa/competitions/ai4d-african-language-dataset-challenge

Figure 3: AI4D Fellowship to develop datasets for Low Resource African Languages

Therefore, this challenge's objective was the creation, curation and collation of good quality African language datasets for a specific NLP task. This task-specific NLP dataset will serve as the downstream task we can evaluate future language models on. This challenge is undergoing and sponsored by GIZ and UNESCO with IDRC support and is hosted in partnership with the Artificial Intelligence for Development Africa (AI4D-Africa) Network.

In early June we have contacted the authors of the winning submissions, to inform them that through the additional support of UNESCO, we are able to work with up-to 5 teams that had outstanding submissions for a further period of few months. During this time, we proposed to support them to further build and annotate the dataset to meet some minimum requirements that we set collaboratively in order to obtain datasets that we can in future use to host shared tasks/ML challenges. This interaction would kick off with a one-day virtual workshop where we set minimum deliverables, agreed on an accountability structure for the coming months and identify what mentorship they may need and set about providing them with it. Data will be published on Zenodo, which is a simple and innovative service enabling researchers to share and showcase research results from all fields of science.

African languages face some challenges that hinder their inclusion on digital platforms, and particularly in academic research. Some of the challenges for the development of NLP for African languages identified by researchers in Africa that this fellowship tackles include[10]:

- Low availability of resources (input data) for African languages that hinders the ability for researchers to do machine translation;
- Lack of benchmarks: Due to the low discoverability and the lack of research in the field, there are no publicly available benchmarks or leader boards to compare machine translation techniques to.

We tackle the challenge of availability to resources by funding the creation of language datasets for African languages. Previous iterations of this work involved an open call for dataset submissions with prize money to incentivize participation. Among our learnings, we found that:

- The challenge framing allowed for anyone to participate. While useful as an exercise in evaluating the interest in such a challenge, the top evaluated submissions came from teams who had been exposed to NLP research work. Targeting such a challenge to NLP researchers could lead to higher quality submissions in future.
- Since the challenge was evaluated monthly, we often received disparate submissions from the same teams as they managed to obtain more data. Instead, one large dataset built over a couple of months would have been the ideal outcome, so in future we'd select and support teams for a sustained period of time to enable them build sizeable datasets.
- Teams composed of individuals from relevant multi-disciplinary backgrounds, including computer scientists, professional translators and linguists, were able to create and annotate datasets that captured fundamental lexical and semantic nuances of languages.

[10] A Focus on Neural Machine Translation for African Languages https://arxiv.org/abs/1906.05685

SIZE OF DATASET ACROSS ALL LANGUAGES: ALMOST 250.000 INSTANCES

Language	Countries	Speakers	Researcher(s)
WOLOF	Senegal,Gambia,Mauritania	10 million	Baamtu Datamation - *Thierno Diop*
EWE	Ghana,Togo	4.5 million	Takwimu Lab - *Kevin Degila* - *Mamboladji Balogoun* - *Godson Kalipe* - *Jamiil Toure*
FONGBE	Benin,Nigeria,Togo	4.1 million	Takwimu Lab - *Kevin Degila* - *Mamboladji Balogoun* - *Godson Kalipe* - *Jamiil Toure*
YORUBA	Nigeria,Benin,Togo,Ghana, Côte d'Ivoire,Sierra Leone	40 million	David Adelani
TUNISIAN ARABIZI	Tunisia	N/A	iCompass Technology - *Chayma Fourati* - *Hatem Haddad* - *Malek Naski*
KISWAHILI	Tanzania,Kenya,Uganda,Rwanda, Burundi,some parts of Malawi,Somalia,Zambia,Mozambique,Democratic Republic of the Congo	100 - 150 million	Davis David
CHICHEWA	Malawi,Mozambique,Zambia, Zimbabwe,South Africa	12 million	Amelia Taylor
TWI	Ghana, Cote d'Ivoire	~20 million	NLP Ghana - *Paul Azunre* - *Lawrence Adu Gyamfi* - *Esther Appiah* - *Felix Akwerh* - *Salomey Osei* - *Samuel Owusu* - *Cynthia Amoaba* - *Salomey Afua Addo* - *Edwin Buabeng-Munkoh* - *Nana Boateng*
LUGANDA	Uganda	8.5 million	Joyce Nakatumba-Nabende Andrew Katumba Jonathan Mukiibi Claire Babirye
9 LANGUAGES	22 COUNTRIES	150 MILLION SPEAKERS	29 RESEARCHERS

Figure 4: Complete data for all languages

With these factors in mind, we designed the fellowship and selected nine of the teams with outstanding submissions to participate. We supported the research teams in the following ways:

- Providing a research stipend to enable them focus on their work and engage specialized expertise, such as translators and linguists, in the creation and annotation of high-quality datasets.
- Creating a platform via which they could engage with each other and other NLP researchers on challenges and technical issues that arose over the duration of the fellowship. This was a bi-monthly call where teams provided updates and tabled challenges they faced and ideas they had for wider discussion.
- Availing access to legal practitioners who advised on matters Intellectual Property, Copyright, Data Protection and Privacy.

Researchers then went ahead creating, curating and annotating datasets for the following NLP tasks, each of which is further discussed below;

- Machine Translation
- Text-to-Speech
- Sentiment Analysis
- Document Classification

While these datasets are not big enough to train models for production-level applications, they will serve to create benchmarks against which subsequent research work can be compared to, thus directly tackling the challenge of a lack of benchmarks for African languages. In

collaboration with Zindi[11], we are hosting Machine Learning challenges on their platform via which we wish to engage the wider African NLP community in this work. Such a challenge is a great opportunity to make widely known the work that has been done in the creation of the dataset and catalyze the exploitation of these datasets. It also incentivizes researchers to experiment with a variety of NLP techniques and compete to see which ones lead to the best model performance.

Machine Translation: Machine Translation, sometimes referred to by the abbreviation MT, is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another. MT products such as Google Translate have become widely used, particularly when travelling abroad. One no longer needs to buy phrase books to learn commonly used phrases in foreign languages.

Within the African NLP research community, MT emerged as the initial focus of the Masakhane movement. The availability of the jw300 dataset and the inclusion of 561 African languages made it such that, with a standard notebook, researchers of all levels could train MT models for the languages they speak and care about by simply editing, within the notebook, the iso-code of the language. This work was seen as merely a starting point for several reasons;

- The material from the jw.org website, which is the original source of the jw300 data is entirely biblical and religious content. This domain bias is apparent in the translation output of the trained models with instances where, for example Canada is translated to Canaan.
- The data is still quite small and not enough for training reliable, production level systems

A natural next step for researcher interested in the task of MT, beyond training an initial baseline model using the data made available through jw300, is to begin sourcing supplementary datasets to begin to balance out the effect of the religious data and to add to the overall dataset with the aim to improve overall translation quality.

Through this fellowship, 5 teams embarked upon the task of creating MT datasets for a total of 15 languages. The sizes of datasets vary, influenced by the time each team had and the methodology selected. We allowed absolute freedom over methodology as the contexts, team composition and partners varied in each case. Below we provide some detail of each dataset.

Language 1: French-Ewe Language MT Dataset

The researchers got in contact with writers and authors who maintain blogs and produce content in the Ewe language. With the permission of these content creators, they collected articles from several websites. Two partner organizations in this work are Togophonie[12], the first website dedicated to learning the most spoken Togolese languages and the Association Wycliffe Togo Linguistic Development Mission[13] dedicated to the translation of the Bible, Literacy, Promotion of Scriptures and Mother Tongues. These two organizations have provided the capacity to translate the datasets sourced from various websites. The dataset has 50,000 instances.

[11] African data science crowdsourcing platform https://zindi.africa/
[12] Togophonie.com http://togophonie.com/
[13] Association Wycliffe Togo Linguistic Development Mission http://www.wycliffetogo.org/

Language Profile: EWE

Dataset
Ewe-French Machine Translation
50,000 instances

Countries where spoken
Ghana,Togo

Number of Speakers
4.5 million

Language family
Niger-Congo, Bantu language

Researcher(s)
Takwimu Lab
- *Kevin Degila*
- *Momboladji Balogoun*
- *Godson Kalipe*
- *Jamiil Toure*

Figure 5: Language profile for Ewe

Language 2: French-Fongbe Language MT Dataset

This dataset was sourced from beninlanguages.com, a cultural project whose objective is to safeguard Beninese cultural heritage while offering a platform to make this culture available through the ages. The dataset has 50,000 instances.

Language Profile: FONGBE

Dataset
Fongbe-French Machine Translation
50,000 instances

Countries where spoken
Benin,Nigeria,Togo

Number of Speakers
4.1 million

Language family
Niger-Congo, Bantu language

Researcher(s)
Takwimu Lab
- *Kevin Degila*
- *Momboladji Balogoun*
- *Godson Kalipe*
- *Jamiil Toure*

Figure 6: Language profile for Fongbe

Language 3: Yoruba Language Dataset

This dataset has been sourced from several online sources, among them the jw300 dataset, a twitter profile that publishes Yoruba proverbs and GlobalVoices.org. Working with a team of professional translators, volunteers on the Yoruba team of GlobalVoices.org, as well as linguists, the team has translated content obtained from web sources as well as done diacritic verification to make sure that the Yoruba accents are used correctly.

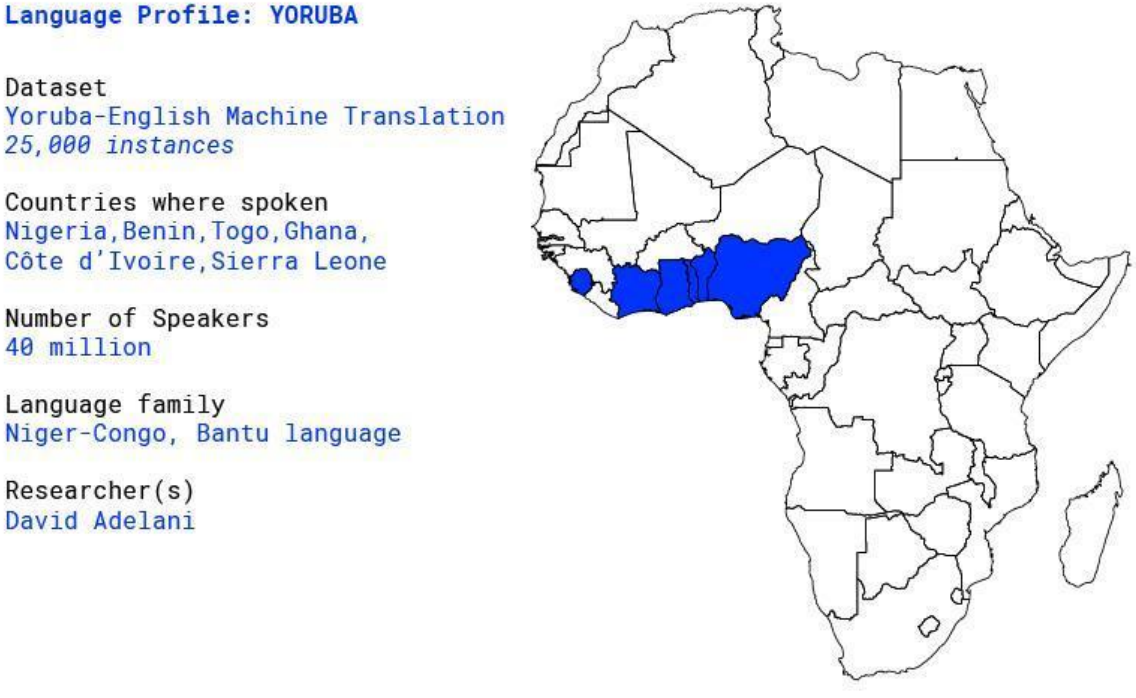

Figure 7: Language profile for Yoruba

Language 4: Luganda Language MT Dataset

Makerere University has been able to leverage translation work done by the Department of African Languages in translating several texts from English to Luganda, most notably the novel Animal Farm by George Orwell, made freely available from Project Gutenberg[14]. In addition to this work, they are working to translate transcripts of plays and African Story books with the Department of African Languages.

With this content, the NLP researchers have worked to align these translations at sentence level and prepare them for the task of Machine translation. This has involved deliberation on how to handle translations that are not necessarily aligned at sentence level, particularly in instances where the material outputted was an interpretation of the original text rather than a direct translation of it. Particularly with Animal Farm, the originally translated text was intended to be an interpretation of the original text done by a lawyer for the purposes of exploring and capturing the themes in Luganda.

[14] Project Gutenberg of Australia eBooks http://gutenberg.net.au/ebooks01/0100011h.html

Language Profile: LUGANDA

Dataset
2000 voice recordings of Agricultural keywords, 3000 voice recordings of keywords from 100 unique contributors, 50,000 Luganda- English Parallel Sentences categorised around 6 topics

Countries where spoken
Uganda

Number of Speakers
8.5 million

Language family
Niger–Congo
Atlantic–Congo
Volta-Congo
Benue–Congo
Bantoid
Southern Bantoid
Bantu
Northeast Bantu
Great Lakes Bantu
Nyoro-Ganda
Luganda

Researcher(s)
Joyce Nakatumba-Nabende
Andrew Katumba
Jonathan Mukiibi
Claire Babirye

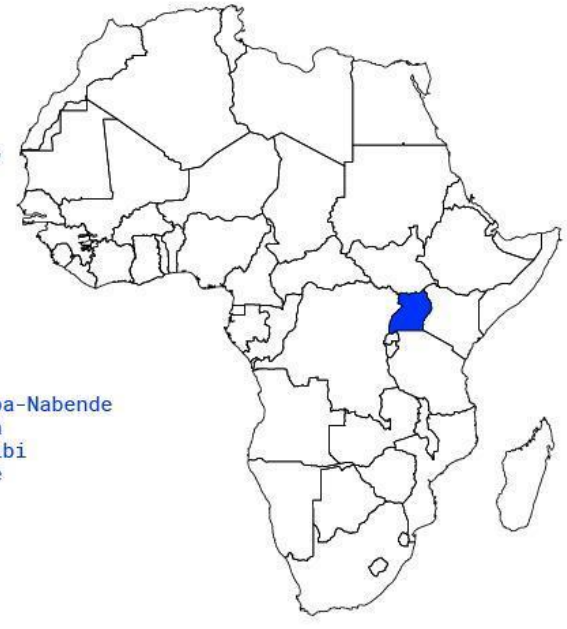

Figure 8: Language profile for Luganda

Language 5: Twi Language Mt Dataset

GhanaNLP[15], having a membership of 80 members signed up, is engaging 10 active members to translate English sentences to Twi. The agreed upon rate is 3 cents per translation with the target being 50,000 translations.

Language Profile: TWI

Dataset
English - (Asante)Twi Machine Translation
50,000 instances

Countries where spoken
Ghana, Cote d'Ivoire

Number of Speakers
~20 million

Language family
Niger-Congo

Researcher(s)
NLP Ghana
- *Paul Azunre*
- *Lawrence Adu Gyamfi*
- *Esther Appiah*
- *Felix Akwerh*
- *Salomey Osei*
- *Samuel Owusu*
- *Cynthia Amoaba*
- *Salomey Afua Addo*
- *Edwin Buabeng-Munkoh*
- *Nana Boateng*

Figure 9: Language profile for Twi

[15] Ghana Natural Language Processing (NLP) https://ghananlp.org/ and https://github.com/GhanaNLP

Language 7: The 11 Official Languages to South Africa

Working with a professional translator, who is part of the team, the researchers have access to the South African Parliament's internal Content Management System (CMS) from which they will align 22,000 sentences from the various languages, as available. These are typically not equally available as they are dependent on the availability of translators. The parliamentary documents being processed are order papers automatic alignment methods will be trialed to potentially speed up the process. Outside of building these resources purely for academic research purposes, areas of evaluating MT tools and their application in the real world are now actively being explored.

Speech Processing

Speech processing technologies, i.e. Automatic Speech Recognition and Text-to-Speech technologies are increasingly being perceived as having potential great wins for African markets. Innovation in this area would enable opening up of access to not only populations that are locked out of digital platforms due to lack of literacy, but with the building of these tools with a focus on multilingualism and inclusion of African languages, they provide the potential to include even those only able to communicate in their mother tongue. A text-to-speech (TTS) system converts normal language text into speech while a Speech Recognition is the ability of a machine to identify words spoken aloud and convert them into readable text.

Language 8: Wolof Language Dataset

Through the dataset challenge that preceded this fellowship, a team from Senegal submitted a Wolof dataset crowdsourced and annotated for the task of speech recognition. Among their motivations was the fact that 50% of the population in Senegal is illiterate and while ASR tools have become prevalent in the past couple of years, they are largely focused on foreign languages. The dataset created is specifically for the transport domain. In collaboration with a local startup, Weego, which is a collaborative transit app that aims to make travel by public transportation easier.

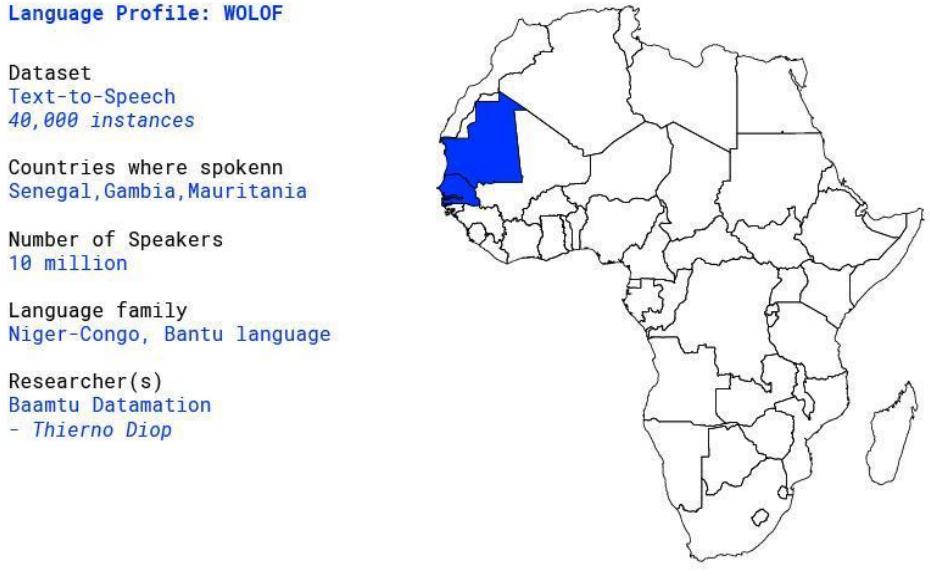

Figure 10: Language profile for Wolof

Through this Fellowship, the team chose to expand this work by building a Wolof TTS dataset containing 40,000 instances. Funding from the fellowship enabled the purchase of high-quality recording equipment and has been used to cover the cost of voice actors. This work has also involved creating a dataset of sentences that adequately cover the phonetic diversity within the Wolof language. These sentences and phrases are the material that will be recorded by the voice actors. With 2 actors, one male and one female, each is recording 20,000 utterances.

Education emerges as another application area where voice technologies can be leveraged in African contexts. An initial focus within the Masakhane community is the use of voice technologies in the context of childhood curricula; ie. literacy education, primary mathematics and primary science curricula.

In low literacy, rural settings, it has been found that a learner's inability to adequately communicate in English, French or Portuguese (colonial languages that have become official languages) contributes to their poor performance in schools. The development of Intelligent Tutoring Systems, with a conversational agent as the main component, would provide additional material and support beyond what can be provided by a teacher within a classroom setting.

Sentiment Analysis

Sentiment analysis (also known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

This technology is used in many social networking services, e-commerce websites and more broadly on platforms where users can provide text reviews, comments or feedback on items. This user generated text is a rich source of user's sentiment about items. Depending on the use case, these mined opinions can discern market or audience sentiment for product and content development.

Language 9: Tunisian Arabizi Language Dataset

This Fellowship is supporting the development of a sentiment Analysis dataset for Tunisian Arabizi, which is an Arabic dialect increasingly gaining use in Tunisia. This dataset contains 100,000 instances. These are obtained by collecting comments from Youtube and Instagram. These are collected from Tunisian content and the comments parsed to ensure they are indeed Tunisian Arabizi and not another language. An initial batch of comments were manually labeled and then subsequently, to hasten the process, a classification model has been trained to do the annotation. The team however still verifies the labels on each comment.

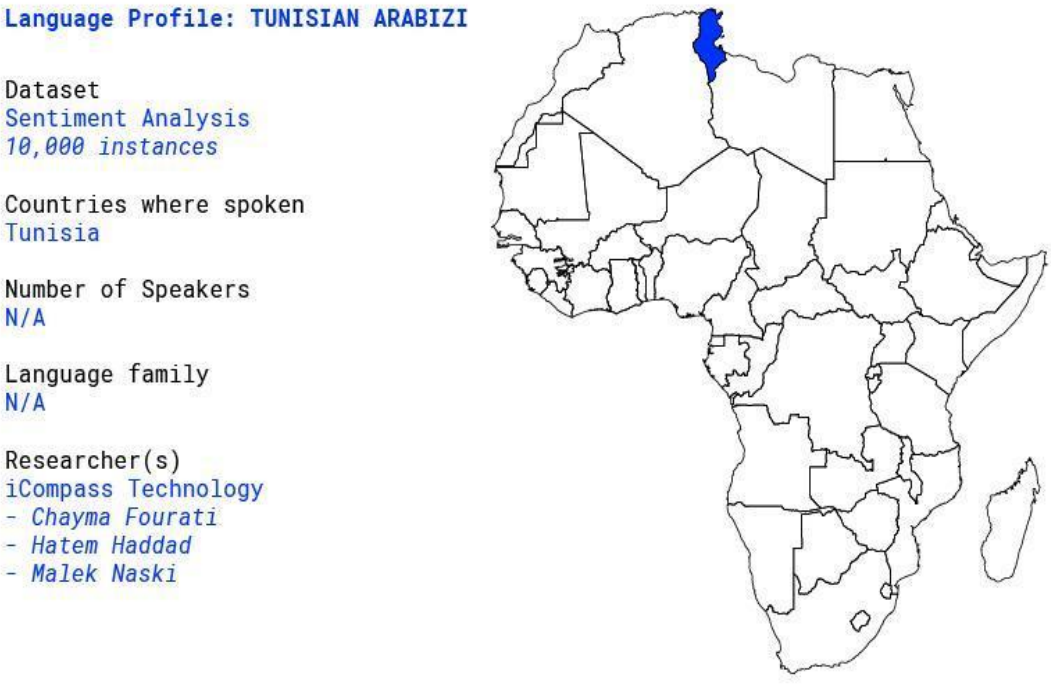

Figure 11: Language profile for Tunisian Arabizi

While sentiment analysis has found great application in business settings as a standalone component, the technique is also being employed as a part of fake news detection systems. With the rapid production of fresh news content and their proliferation on various news and social media platforms, it is becoming increasingly important in today's society to be able to detect fake news.

Equally pertinent is the detection and curbing of hate speech, another use case that is reliant on Sentiment Analysis. Online hate speech is the expression of conflicts between different groups within and across societies, attacking a person or a group based on their race, religion, ethnic origin, sexual orientation, disability, or gender.

Document Classification

Document Classification is a task to assign a document to one or more classes or categories. This may be done manually or algorithmically. The documents to be classified may be texts, images, music, etc. Each kind of document possesses its special classification problems. When not otherwise specified, text classification is implied. A domain area for Document Classification problems is in classifying news articles, particularly online, into their relevant categories such as politics, sports, weather, finance, etc.

Language 10: Chichewa And Kiswahili Language Dataset

This Fellowship is supporting the development of two document classification datasets. The first for Chichewa, containing 5,000 instances and the second for Kiswahili containing 10,000 instances. Both datasets have been obtained by scraping data from several newspaper websites identified to publish data in the two languages.

Language Profile: KISWAHILI

Dataset
Document Classification
10,000 instances

Countries where spoken
Tanzania, Kenya, Uganda, Rwanda, Burundi, some parts of Malawi, Somalia, Zambia, Mozambique, Democratic Republic of the Congo (DRC)

Number of Speakers
100 - 150 million

Language family
Niger-Congo, Bantu language

Researcher(s)
Davis David

Figure 12: Language profile for Kiswahili

The Chichewa data is being manually annotated with the category of the news item while the Kiswahili data is being scraped per category, which then makes the process of annotation easier and faster. Further to news classification, the automatic classification of documents on the internet, particularly those pertaining to information and knowledge is a useful feature that would enable codification of material as it is made available in African languages and on different media.

Language Profile: CHICHEWA

Dataset
Document Classification
5000 instances

Countries where spokenn
Malawi,Mozambique,Zambia, Zimbabwe,South Africa

Number of Speakers
12 million

Language family
Niger-Congo, Bantu language

Researcher(s)
Amelia Taylor

Figure 13: Language profile for Chichewa

Result 3: Machine Learning Data Challenge Series

These AI4D datasets are further made into competitions of five NLP challenges hosted on Zindi as part of this AI4D's ongoing African language NLP project, which are a continuation of the African language dataset challenges we hosted in 2020[16]. The competition engagement in the challenges shows a total of 153,926 unique page views across 111 countries for the implementation of the challenges: Tunizi Arabizi, Yoruba, Ewe & Fongbe, Chichewa. Wolof went live on 12 Feb 2020 so statistics are not available yet. The current statistics of each challenge are the following:

- AI4D iCompass Social Media Sentiment Analysis for Tunisian Arabizi (20 November 2020—29 March 2021): 539 Data scientists enrolled, 213 Data scientists on the leaderboard, 3 630 Submissions, Accuracy score: 0.94
- AI4D Malawi News Classification Challenge (22 January—10 May): 218 Data scientists enrolled, 69 Data scientists on the leaderboard, 686 Submissions, Accuracy score: 0.64
- AI4D Takwimu Lab – Machine Translation Challenge (18 December 2020—26 April 2021): 134 Data scientists enrolled, 11 Data scientists on the leaderboard, 142 Submissions, BLEU score: 0.35
- AI4D Yorùbá Machine Translation Challenge (4 December 2020—12 April 2021): 314 Data scientists enrolled, 33 Data scientists on the leaderboard, 285 Submissions, BLEU score: 0.43
- AI4D Baamtu Datamation - Automatic Speech Recognition in WOLOF (12 February—24 May)

Challenge 1: Automatic Speech Recognition in Wolof

Basic description: In a country such as Senegal, where about 50% of the population is illiterate, technologies and applications that are designed to be used by people who can read are not as effective as they could be. In this competition, the aim is to use Automatic Speech Recognition (ASR) techniques in the Wolof language to help illiterate people to interact with apps with just their voice, in a language they can already speak.

Challenge description: The challenge[17] is focused on a public transport use case for two reasons. First, many users of public transport can't read or speak French, so they can't interact with existing apps that help passengers to find a bus for a given destination. And second, there is already an existing app in Senegal, WeeGo, which helps passengers to get transport information. The goal of this competition is to build an ASR model that will help illiterate people use existing apps to find which bus they can take to reach their destination, without having to know how to read or write.

Challenge 2: Malawi News Classification Challenge

Basic description: Algorithms for text classification still contain some open problems for example dealing with long pieces of texts and with texts in under-resourced languages. This challenge gives participants the opportunity to improve on text classification techniques and algorithms for text in Chichewa. The texts are of varying length, some being quite long and will pose some challenges in chunking and classification. The texts are made up of news articles.

[16] Zindi and AI4D build language datasets for African NLP https://zindi.medium.com/zindi-and-ai4d-build-language-datasets-for-african-nlp-34a4d0ea129
[17] AI4D Baamtu Datamation - Automatic Speech Recognition in WOLOF https://zindi.africa/competitions/ai4d-baamtu-datamation-automatic-speech-recognition-in-wolof

Challenge description: The objective of this challenge[18] is to classify news articles. We hope that the solutions will illustrate some challenges and offer solutions. Algorithms for text classification have come a long way, but classifying long texts and working with under-resourced languages can still pose difficulties. This challenge gives participants the opportunity to improve on text classification techniques and algorithms for text in Chichewa. The texts are made up of news articles or varying lengths. The objective of this challenge is to classify these articles by topic. We hope that the solutions will illustrate some challenges and offer solutions. Chichewa is a Bantu language spoken in much of Southern, Southeast and East Africa, namely the countries of Malawi and Zambia, where it is an official language, and Mozambique and Zimbabwe where it is a recognized minority language.

Challenge 3: Takwimu Lab - Machine Translation Challenge

Basic description: Ewe and Fongbe are Niger–Congo languages, part of a cluster of related languages commonly called Gbe. Fongbe is the major Gbe language of Benin with approximately 4.1 million speakers, while Ewe is spoken in Togo and southeastern Ghana by approximately 4.5 million people as a first language and by a million others as a second language. They are closely related tonal languages, and both contain diacritics that can make them difficult to study, understand, and translate.

Although those languages are at the core of the economic and social life of at least 3 major West African capital cities (namely Cotonou, Lome and Accra), they are today mostly spoken and very rarely written. Due to that fact (among other reasons), there is very little official or formal communication in those languages, leaving non-French/English speakers often unable to access critical facilities like education, banking, and healthcare. This challenge is part of an initiative that wishes to bring down the barriers between African local language speakers and modern society.

Challenge description: The objective of this challenge[19] is to create a machine translation system capable of converting text from French into Fongbe or Ewe. Applicants may train one model per language or create a single model for both and not use any external data, so a key component of this competition is finding a way to work with the available data efficiently. This is a pioneer competition as far as low-resourced West African languages are concerned. A good solution would be a model that can be improved upon or used by researchers across the world to create APIs that can be integrated into day-to-day tools like ATMs, delivery applications etc., and help bridge the gap between rural West Africa and the modernized services.

Challenge 4: Yorùbá Machine Translation Challenge

Basic description: Machine translation (MT) is a popular Natural Language Processing (NLP) task which involves the automatic translation of sentences from a source language to a target language. Machine translation models are very sensitive to the domain they were trained on which limit their generalization to multiple domains of interest like legal or medical domains. The problem is more severe in low-resource languages like Yorùbá where the most available datasets used for training are in the religious domain like JW300. How can we train MT models to generalize to multiple domains or quickly adapt to new domains of interest? In this challenge, you are provided with 10,000 Yorùbá to English parallel sentences sourced from multiple domains like news articles, ted talks, movie transcripts, radio transcripts, software localization texts, and other short articles curated from the web. Your task is to train a multi-domain MT model that will perform very well for practical use cases.

[18] AI4D Malawi News Classification Challenge https://zindi.africa/competitions/ai4d-malawi-news-classification-challenge
[19] AI4D Takwimu Lab - Machine Translation Challenge https://zindi.africa/competitions/ai4d-takwimu-lab-machine-translation-challenge

Challenge description: The goal of this challenge[20] is to build a machine translation model to translate sentences from Yorùbá language to English language in several domains like news articles, daily conversations, spoken dialog transcripts and books. The applicant's solution will be judged by how well the translation prediction is semantically similar to the reference translation. The translation models developed will assist human translators in their jobs, help English speakers to have better communication with native speakers of Yorùbá, and improve the automatic translation of Yorùbá web pages to English language.

Challenge 5: Social Media Sentiment Analysis for Tunisian Arabizi

Basic description: On social media, Arabic speakers tend to express themselves in their own local dialect. To do so, Tunisians use 'Tunisian Arabizi', where the Latin alphabet is supplemented with numbers. However, annotated datasets for Arabizi are limited; in fact, this challenge uses the only known Tunisian Arabizi dataset in existence. Sentiment analysis relies on multiple word senses and cultural knowledge, and can be influenced by age, gender and socio-economic status. For this task, we have collected and annotated sentences from different social media platforms.

Challenge description: The objective[21] of this challenge is to, given a sentence, classify whether the sentence is of positive, negative, or neutral sentiment. For messages conveying both a positive and negative sentiment, whichever is the stronger sentiment should be chosen. Predict if the text would be considered positive, negative, or neutral (for an average user). This is a binary task. Such solutions could be used by banking, insurance companies, or social media influencers to better understand and interpret a product's audience and their reactions.

Result 4: Additional projects

The AI4D projects gave us exposure across African AI/ML communities in pursuing additional funding with the Lacuna Fund in support of Masakhane, the open research, participatory, grassroots NLP initiative for Africans by Africans, with the aim of putting African research in NLP on the map, by holistically tackling the problems facing NLP. Lacuna Fund is the world's first collaborative effort to provide data scientists, researchers, and social entrepreneurs in low- and middle-income contexts globally.

Projects: In the first cohort of supported projects we have:

- **South Africa**: Build a multilingual parallel corpus of African research, by translating African Masakhane MT: Decolonizing Scientific Writing for Africa
- **Uganda**: Masakhane NER: Named Entity Recognition & Parts of Speech datasets for African Languages

In addition, some areas identified as key to include for future directions of this work included:

- Capacity Building: Support the creation of a formal "Masakhane[22] - African NLP Researcher Collaborative Network" based on the results of the individual participants and projects and for the fellowship as a whole
- Technology: Build a "Masakhane Web TTS platform for Speech-to-Text conversion" to accurately convert speech into text using an API powered by AI4D's funded AI technologies. Beginning with a case study with Wolof language and to be extended to any African language and in line with "Masakhane MT platform for Machine Translation".

[20] AI4D Yorùbá Machine Translation Challenge https://zindi.africa/competitions/ai4d-yoruba-machine-translation-challenge
[21] AI4D iCompass Social Media Sentiment Analysis for Tunisian Arabizi https://zindi.africa/competitions/ai4d-icompass-social-media-sentiment-analysis-for-tunisian-arabizi
[22] Masakhane - A grassroots NLP community for Africa, by Africans https://www.masakhane.io/

5. Project outputs

The following activities were supported by the project during the entire reporting period (M1-M15), including the originally planned project objectives and additional ones:

Outcome 1: Call for proposals for applications

The AI4D Africa call for applications has generated 11 projects. The projects piloted a mentorship programme partially inspiring the Deep Learning Indaba[23] programme with procedures for monitoring the project progresses, as well as timelines and deliverables. We also released a public form[24] for researchers to give feedback on their exploitation plans. Datasets were uploaded to Zenodo.

Outcome 2: Data creation and data Challenge

The objective of having a portfolio of challenges focused on the creation, curation and collation of good quality African language datasets for a specific NLP task was achieved. This work was taken forward into a more coherent project focused on ***Cracking the Language Barrier for a Multilingual Africa***[25]*.*

Outcome 3: Project website

This result has been achieved[26] and is being redesigned and updated with news, videos, blogs and science talks sections[27].

6. Problems and Challenges

The outreach programme to strengthen the AI4D Network Africa has started on 1 March 2020 and ended on 28 February 2021 after an extension period as on 11 March 2020, the World Health Organization (WHO) officially classified COVID-19 as a pandemic. This impacted all aspects of our project, therefore we asked for an extension of the work on three main project actions namely:

1. Finalize the Fellowship to develop datasets and strengthen capacities and innovation potential for Low Resource African Languages;
2. Finalize the mini-projects and help create effective exploitation routes for post COVID-19 opportunities.

[23] https://deeplearningindaba.com/mentorship/
[24] AI4D mini-projects exploitation plans https://forms.gle/V7fuScsavhF3pMpc9
[25] Fellowship and NLP project https://docs.google.com/document/d/10Zm7AjJCUu-6nkp1qr_BfYFZcJCSnRoVqGhWGsyZYGw/edit?usp=sharing
[26] AI4D website https://ai4d.ai/
[27] AI4D science talks https://ai4d.ai/talks/