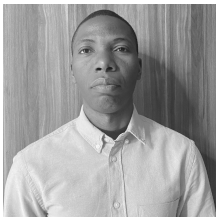
FELERMINO ALI

AI Specialist

felerminoali@gmail.com
+258 846-689-637 / +351 910-51-7193
Maputo, Marracuene
Maputo, Mozambique
felerminoali.github.io/
@felerminoali
felermino-ali
felerminoali



EXPERIENCE

Post-Doc Researcher

Microsoft Research Africa

Oct 2025 - Present — Nairobi, Kenya

- Artificial Intelligence
- Natural Language Processing
- Large Language Modeling
- Speech Recognition

Consultant - NLP & Corpus Development

MAPA Consultores

April 2026 - Present — Remote

- Gates Foundation funded multi-country study (Mozambique, Ghana and South Africa) developing picture-based receptive vocabulary assessments for Grade 3–4 learners in their L1 and in the language of instruction.
- Built L1 and L2 corpora for Mozambican languages from digitised readers, textbooks and children's literature, using automated PDF collection, OCR and web scraping, with a cleaning and text-categorisation pipeline.
- LLM-assisted lemmatisation and morphological analysis of morphologically rich Bantu languages, with expert-in-the-loop review of low-confidence token–lemma mappings.
- Built lemma, instance and source databases (raw and Zipf frequency, part of speech, contextual diversity) and tracked corpus saturation with lexical diversity metrics (TTR, MTLD, HD-D).
- Supported test construction by sampling target words and distractors across frequency bands, parts of speech and domains.

Member of the United Nation Working Group of Tech for Victims of Terrorism Multicultural Solutions

United Nation

April 2026 - Present

- Design policies for development of AI Tech for Victims of Terrorism Multicultural Solutions

Researcher

Laboratory of Artificial Intelligence and Computer Science (LIACC) - Faculty of Engineering of Porto University

Sept 2021 - 2026 — Porto, Portugal

- Researcher and PhD student: Natural Language Processing and Machine Translation Research

Consultant - Software Engineer

Cedsif, IT

Jul 2021 — Maputo, Mozambique

HONOR AWARDS

Valedictorian
Given at graduation ceremony by the former President of Mozambique (His Excellency Mr Armando Emilio Guebuza)

FCT PhD scholarships
shortlisted as first place at "Consórcio de escolas de engenharias" of Science and Technology Foundation - PhD scholarships

SKILLS

Hard-working | Eye for detail | Motivator & Leader

C++ | Python | R | NLP | Machine Translation | Information Retrieval | Data Science | Artificial Intelligence | Statistical Analysis | Agile Management | Software Engineering | JAVA | Web Development

LANGUAGES

Portuguese ●●●●●

English ●●●●◐

Emakhuwa ●●○○○

- Center for the Development of Finance Information Systems (CEDSIF) - CEDSIF is an institution linked to the Ministry of Economy and Finance which was set up to modernize the public finance management information systems for all state and municipal bodies
- Full stack developer for Government Municipality Management System.

Consultant - Moodle Integration - Developer

Mwetana

Jan 2020 - Jun 2021 | Remote - Monrovia, Liberia

- Integrate Moodle Learning Management System with SchooLink™ SchoolManagement System (http://moodle.schoolinkcx.com/)

Consultant - Software Engineer

SoliderMed

Jan 2019 - Jun 2021 | Pemba - Cabo Delgado, Mozambique

- Software development of an AI-based alert system based on labor time to prevent maternal and neonatal morbidity and mortality in Mozambique.

Consultant - Library Management System Iris University

IRIS UNIVERSITY

Jun 2020 | Pemba, Mozambique

- Deployment of Iris University library management system
- http://koha.universidadeiris.com/
- Training

Consultant - eLearning Training

INSTITUTO DE SAÚDE DE PEMBA

2020 | Pemba, Mozambique

- Students and Instructors Training on Moodle e-learning Platform

Consultant - Moodle deployment and Training

Unirovuma University

2021 | Nampula, Mozambique

- Students and Instructors Training on Moodle Platform

Administration | Faculty Member | AI Researcher and Software Developer

Lurio University

Pemba, Mozambique

- Assistant Lecturer of the following subjects: Object Oriented Programming, Data Structure, Database Lab, Software Engineering Lab, Artificial Intelligence, Multimedia, Programming Foundation
- Developer of E-Sira (seul.unilurio.ac.mz)
- Deputy Director of Faculty of Engineering
- Research

Research Assistant

Research Assistant

SEGi University, Malaysia

EDUCATION

Ph.D. in Informatics Engineering

Faculty of Engineering of the University of Porto

Sept 2021 – 2026 | Porto, Portugal

- Thesis: Low-Resource Machine Translation for Emakhuwa: Transfer Learning, Data Augmentation, and Lexical Resource Integration
- Link: `https://scholar.google.com/citations?view_op=view_citation&hl=pt-PT&user=YJsI-cOAAAAJ&sortby=pubdate&citation_for_view=YJsI-cOAAAAJ:9ZlFYXVOiuMC`

M.Sc. in Information Technology

SEGi University

Sept 2014 – June 201 | Kuala Lumpur, Malaysia

B.Sc. in Computer Engineering

Lurio University

Sept 2008 – June 201 | Pemba, Mozambique

EXHIBITIONS

Exhibitor at the 2025 Osaka Expo, representing Mozambique with "*Ntxuva Inteligente*"—an AI-powered interpretation of the traditional Mozambican game, Ntxuva.

- Reasearch on Data Mining applications on E-learning systems

PROJECTS

AfriSenti: African Languages Sentiment Analysis Datasets and Benchmarks

Role: Organizer

- Creation of sentiment-analysis datasets and benchmarks for African languages, helping researchers build and evaluate NLP models that can understand opinions and sentiment in languages that are poorly represented in mainstream AI datasets.
- Winner of best paper award at the African Languages NLP Workshop 2024

Text Based Emotion Analysis (SemEval 2025 Task)

Role: Organizer for Emakhuwa Language

- Task: Determining what emotion most people will think the speaker may be feeling given a sentence or short text snippet uttered by the speaker.

Emakhuwa Machine Translation Dataset Collection

Lacuna Fund/Meridian Institute

2022-2024

- Awarded a grant to create the first large-scale dataset for Machine Translation of the Emakhuwa language, Mozambique's widely spoken indigenous language.
- https://lacunafund.org/2021-african-language-awardees/
- aprox. 50k USD

MOZNLP-UP: Scalling Up Mozambican Languages

Lacuna Fund/Meridian Institute

2025-Present

- Awarded a grant to create the expand the first lacuna fund project for Xichangana, Nyanja and Sena.
- aprox. 100k USD

Meta, Common Voice Data Collection

Mozilla

2025

- Collection of audio recordings in two Mozambican languages, Nyugwe and Ekoti, for speech-to-text and text-to-speech AI applications.

Google exploreCSR

Google

1 yr

Promoting computing research among underprivileged students in the North of Mozambique. The project focused on boosting self-efficacy, fostering belonging, providing practical skills in Computer Sciences, and encouraging graduate studies.
https://research.google/outreach/explore-csr/recipients/

CATfrica - Project

Personal Project

First Computer Assisted Translation (CAT) tool for for Mozambican languages https://mozcat.vercel.app/

Models and Data Repos

Hugging Faces

- Contributor and Member of LIACC HF Repo: https://huggingface.co/LIACC

- Personal Repo: https://huggingface.co/felerminoali

TRAINING

11th INTERNATIONAL SCHOOL ON DEEP LEARNING

https://irdta.eu/

July 2024 — Maia, Portugal

- Summer School: 6-day event (15-19 July, 2024) cover multiple topics including: computer vision, neurosciences, speech recognition, language processing, human-computer interaction, drug discovery, health informatics, medical image analysis, recommender systems, advertising, fraud detection, robotics, games, finance, biotechnology, physics experiments, biometrics, communications, climate sciences, geographic information systems, signal processing, genomics, materials design, video technology, social systems, etc.

MEDCIDS - Winter School 2024

Faculty of Medicine of the University of Porto

February, 2024 — Porto, Portugal

- Winter School: 3-day event (5-7 February 2024) Reliability and agreement studies

The 12th Lisbon Machine Learning

Instituto Superior Técnico, Instituto de Telecomunicações

July, 2022 — Lisbon, Portugal

- Summer School: 6-day event (24-29 July 2022) covering a range of machine learning topics, from theory to practice, that are important in solving natural language processing problems arising in different application areas.

NVIDIA DLI Certificate

June, 2022

- Fundamentals of Deep Learning

Microsoft Natural Language Processing Processing Event

April, 2022

- Microsoft Learn Student Ambassador

ICMC 50: Python for Natural Language Processing

Institute of Mathematics Sciences and Computation

2021 — Brazil

- Python for Natural Language Processing

Business Data Science Summer School

Schools of Economics and Business of Erasmus University Rotterdam, University of Amsterdam and Vrije Universiteit Amsterdam

August 17-21, 2020 — Amsterdam, Netherlands

- Deep Learning

Professional Scrum Master

2015

- Training on PSM I certification by Joshua Partogi

Cisco Certified Network Associate

Nov 2014

- CCNA certification

CONFERENCES AND WORKSHOPS

African NLP

2025 — Austria, Vienna

- **Paper title**: **MOZ-Smishing: A Benchmark Dataset for Detecting Mobile Money Frauds**

The 2024 Conference on Empirical Methods in Natural Language Processing

Nov. 2024 — Miami, Florida, USA

- **Paper title**: Building Resources for Emakhuwa: Machine Translation and News Classification Benchmarks.

Open Language Data Initiative WMT workshop

Nov. 2024 — Miami, Florida, USA

- **Paper title**: Expanding FLORES+ Benchmark for More Low-Resource Settings: Portuguese-Emakhuwa Machine Translation Evaluation.

First Workshop on NLP for Indigenous Languages of Lusophone Countries

2024 — Galicia

- **Paper title**: Network-based Approach for Stopwords Detection.

Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)

May, 2024 — Torino, Italia

- **Paper title**: Detecting Loanwords in Emakhuwa: An Extremely Low-Resource Bantu Language Exhibiting Significant Borrowing from Portuguese.

LITHME WG1-WG7 Joint Workshop: Bridging the gap between technology and professionals

Oct. 28-30, 2023 — Budapest, Hungary

- **Paper title**: Data Augmentation for Low-Resource Neural Machine Translation Using Loanword Spelling Inconsistencies.

Language in the Human-Machine Era LITHME

University of Groningen, Campus Fryslân

Oct. 15-16,2022 — Leeuwarden, Netherlands

- **title**: Loanword detection in Emakhuwa language.

AfricAI Conference

University of Groningen

Oct. 15-16,2022 — Kigali, Rwanda

- **title**: Lacuna Fund Grantee Convening.

Doctoral Symposium, EPIA 2022

21st EPIA Conference on Artificial Intelligence

Oct. 15-16,2022 — Lisbon, Portugal

- **title**: Machine Translation for Emakhuwa of Mozambique.

AfricaNLP Workshop

2021

- **title**: Towards a parallel corpus of Portuguese and the Bantu language Emakhuwa of Mozambique

INVITED TALKS

JKUAT HCI Workshop

2026 Kenya

- Speaker at panel discussion on Human-Computer Interaction
- https://jkuat-hci-workshop.vercel.app/program

AfricaNLP Workshop

2026 Morroco

- **title**: Beyond Parallel Data: Harnessing External Knowledge for Low-Resource MT
- https://sites.google.com/view/africanlp2026/home

AIMS

2026 Senegal

- **title**: Machine translation for African languages

Workshop on Translation Technology

Center of Linguistics of Porto University

2022 Porto, Portugal

- **title**: Machine Translation and CAT tools

Center of African Studies of University Eduardo Mondlane

2022 Maputo, Mozambique

- **title**: Challenges and Opportunities for Mozambican Languages in the age of Large Language Models

1st Seminar on Security and resiliency of comunications in Mozambique

2022 Maputo, Mozambique

- **title**: Simishing SMS attacks detection on Mobile Money Transfer User
- **Link**: https://www.facebook.com/watch/live/?ref=watch_permalink&v=2780324678938427

IndabaX Mozambique

2021 Pemba, Mozambique

- **title**: Mozambican languages in the Natural Language Processing panorama

II Jornadas de Rádiodifusão (XII Radio Broadcasting Conferences)

2021 Maputo, Mozambique

- **title**: Computational Linguistics and challenges for Mozambican language

REVIEWING AND CONFERENCE ORGANIZATION

- Organizing committee for PROPOR conference 2027
- Participate as Microsoft Sponsor at DeepLearning Indaba 2026 (Nigeria)
- Member of Jury of Jornadas Científicas do Banco de Moçambique, 2026 – **Link**: https://www.bancomoc.mz/pt/media/jornadas-cientificas/
- Active Reviewer for *ACL Rolling Review*
- Organizing committee for AfricanLNLP workshop 2025 – **Link**: https://sites.google.com/view/africanlp2025/home?pli=1&authuser=0
- Organizing committee for Beyond Benchmarks: Scaling Multi-Turn Participatory AI Evaluations in Health and Education – **Link**: https://yux-cultural-ai-lab.github.io/Beyond-Benchmarks-Workshop-DLI-2026/#organizers
- Organizing committee for LUHME workshop 2025 **Link**: https://luhme.up.pt/organizers-and-committee/
- Organizing committee for Jornadas Científicas da FECN - 2018, 2019, 2020

PUBLICATIONS

- **F. D. M. A. Ali**, M. Ochieng, O. Ekwejunor-Etchie, A. Famoti, J. O'Neill, and D. Paul, "Latent Core Tokenizer: Compress, but Meaningfully," *arXiv preprint*, Oct. 10, 2026, Preprint; under review. DOI: 10.48550/arXiv.2610.12376 arXiv: 2610.12376 [cs.CL].
- M. Muchai et al., "How Architecture, Typology, and Adaptation Strategies Shape ASR for Low-Resource African Languages: A Controlled Empirical Study," *arXiv preprint*, Oct. 28, 2026, Preprint. DOI: 10.48550/arXiv.2610.39004 arXiv: 2610.39004 [cs.CL].
- S. H. Muhammad et al., "Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages," *arXiv preprint arXiv:2502.11926*, 2025.
- S. M. Saide, E. L. A. Sarmento, and **F. M. D. A. Ali**, "Cryptojacking malware detection in docker images using supervised machine learning," *International Conference on Intelligent and Innovative Computing Applications*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:257780017
- **F. D. M. A. Ali**, A. Caines, and J. L. A. Malavi, "Towards a parallel corpus of portuguese and the bantu language emakhuwa of mozambique," 2021. arXiv: 2104.05753 [cs.CL].
- **F. D. M. A. Ali**, E. Gimo, and S. M. Saide, "A minimax agent for playing ntxuva game – the mozambican variant of mancala," pp. 1–5, 2020. DOI: 10.1109/icABCD49160.2020.9183848
- B. J. Fonseca, A. Felermino D. M. A., and S. M. Saide, "A deep convolutional neural network for classifying waste containers as full or not full," pp. 54–59, 2019. DOI: 10.1109/ISC246665.2019.9071746

Conference Proceedings

- **F. D. M. A. Ali**, D. L. Mateus, and M. V. Mangue, "Measuring the cost of variety conflation in multilingual mt evaluation: Adding mozambican xichangana, nyanja and sena to flores+," in *Proceedings of the Ninth Conference on Machine Translation*, Budapest, Hungary: Association for Computational Linguistics, Nov. 2026.
- M. Ochieng et al., "Toward cultural alignment: Human-centered evaluation of multimodal ai stories across five african communities," in *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, Budapest, Hungary: Association for Computational Linguistics, Nov. 2026.
- **F. D. M. A. Ali**, R. Sousa-Silva, and H. L. Cardoso, "Leveraging Loanword Constraints for Improving Machine Translation in a Low-Resource Multilingual Context," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.
- S. Li et al., "SSA-COMET: Do LLMs outperform learned metrics in evaluating MT for under-resourced African languages?" In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.
- S. H. Muhammad et al., "SemEval-2025 task 11: Bridging the gap in text-based emotion detection," in *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria: Association for Computational Linguistics, 2025.
- **F. D. M. A. Ali**, G. de Jesus, H. L. Cardoso, S. Nunes, and R. Sousa-Silva, "Network-based approach for stopwords detection," in *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, P. Gamallo et al., Eds., Santiago de Compostela, Galicia/Spain: Association for Computational Lingustics, Mar. 2024, pp. 55–63. [Online]. Available: https://aclanthology.org/2024.propor-2.9
- **F. D. M. A. Ali**, H. L. Cardoso, and R. Sousa-Silva, "Building resources for emakhuwa: Machine translation and news classification benchmark," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024.
- **F. D. M. A. Ali**, H. L. Cardoso, and R. Sousa-Silva, "Expanding flores+ benchmark for more low-resource settings: Portuguese-emakhuwa machine translation evaluation," in *Proceedings of the Ninth Conference on Machine Translation*, Maimi, Florida, USA: The Association for Computational Linguistics, Nov. 2024. [Online]. Available: https://www2.statmt.org/wmt24/pdf/2024.wmt-1.45.pdf
- **F. D. M. Ali**, H. Lopes Cardoso, and R. Sousa-Silva, "Building resources for emakhuwa: Machine translation and news classification benchmarks," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, November: Association for Computational Linguistics, Nov. 2024.
- **F. D. M. Ali**, H. Lopes Cardoso, and R. Sousa-Silva, "Detecting loanwords in emakhuwa: An extremely low-resource Bantu language exhibiting significant borrowing from Portuguese," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 4750–4759. [Online]. Available: https://aclanthology.org/2024.lrec-main.425
- F. M. E. Manuel, S. M. Saide, **F. D. M. A. Ali**, and S. Lotfi, "Ocular cataract identification using deep convolutional neural networks," in *2023 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, 2023, pp. 1–5. DOI: 10.1109/icABCD59051.2023.10220532

- S. Muhammad et al., “AfriSenti: A Twitter sentiment analysis benchmark for African languages,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 13 968–13 981. DOI: 10.18653/v1/2023.emnlp-main.862
- E. Á. Cardoso, **F. D. M. A. Ali**, and S. M. Saide, “Snake species identification using deep convolutional neural networks,” in *2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2022, pp. 0202–0209. DOI: 10.1109/IEMCON56893.2022.9946504