

Local Language Processing in Robots



Dr. Emeka Ogbuju

Department of Computer Science,
Miva Open University, Abuja

3rd Artificial Intelligence & Robotics Summit



Dr. Ogbuju is a recognized Fellow of Scholars Academic and Scientific Society, and Nigeria School on Internet Governance. He is a Professional Member of the Association for Computing Machinery, Pan-African Scientific Research Council, and Nigeria Computer Society. Currently, he serves as the Head of the Department of Computer Science and Acting Dean of the School of Computing at Miva Open University, Abuja.

Theory

Local Nigerian Language Processing in Robots involves developing advanced NLP capabilities for robots to respond in indigenous Nigerian languages. It's an innovative initiative that aims to bridge linguistic gaps and enhance human-robot interaction in local communities. We start by curating language resources for Yoruba, Igbo, or Hausa language with focus on enabling assistive techs in education, healthcare, and everyday tasks tailored to local contexts. Cultural inclusivity is vital in AI systems and this line of evolving application is aimed at facilitating seamless communication and deeper connection between technology and users in Nigeria.

Which is correct?

“Teaching robots to speak our local dialects could make it easier for AI to exploit us”

VS

“Humans tend to trust robots more when they speak our local dialects”

Over 500 Languages in Nigeria

- **Niger-Congo:** Yoruba, Igbo, Edo, Hausa and Fulfulde, Nupe and Gbagyi.
- **Nilo-Saharan:** Kanuri and Teda
- **Languages of the Plateau:** Berom and Tarok
- **Others:** Egbira, Okun, Igala

Nigerian Languages Are Low-Resourced

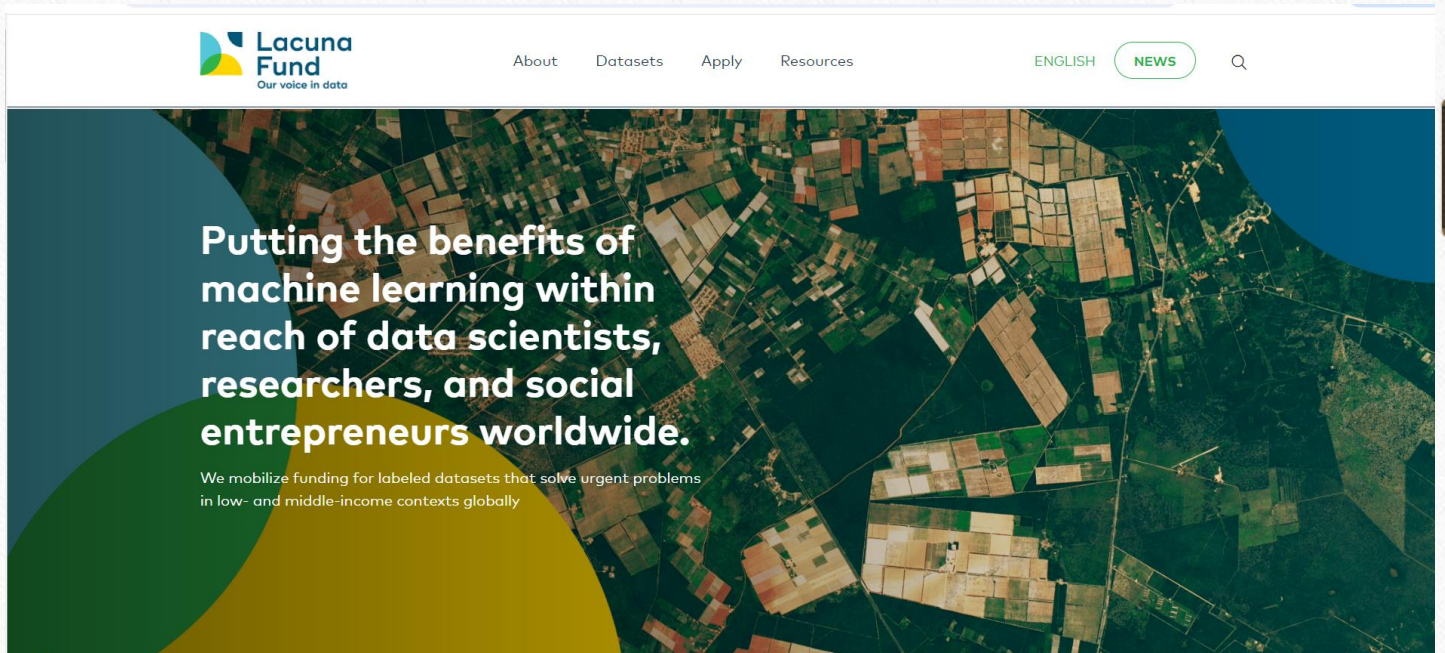
- Limited linguistic data and resources are available for natural language processing (NLP) tasks, such as **machine translation, speech recognition, and named entity recognition**.
- Lack comprehensive digital text corpora, language models, annotated data, and other resources that are essential for training and developing NLP systems.

What is the official language of Nigeria?

- English, French, Portuguese and Spanish are high resourced and most African nations adopt them leaving ours impoverished
- We need to have **indigenous robots**, not colonized robots that speaks English only
- The need to curate local language resources abound

Lacuna Fund

- <https://lacunafund.org/>



Hausa-NLP Open Community

- <https://github.com/hausanlp>

 [Awesome-HausaNLP](#) Public

A curated list of resources dedicated to Natural Language Processing (NLP) in Hausa language

☆ 3 🍴 1

 [HausaVisualGenome](#) Public

 [NaijaSenti](#) Public

Forked from [shmuhammadd/NaijaSenti](#)

This is a repository for NaijaSenti. A Lacuna Funded Project for the development of sentiment corpus for four Nigerian languages: Igbo, Hausa, Yoruba and Pidgin.

☆ 24 🍴 14

 [HERDPhobia](#) Public

A dataset for Hate speech detection against Fulani Herdsmen in Nigeria

☆ 2 🍴 2



Nigerian Sentiment Analysis

<https://github.com/hausanlp/NaijaSenti>



license MIT license CCBY visit our site

NaijaSenti is an open-source sentiment and emotion corpora for four major Nigerian languages. This project was supported by [lacuna-fund](#) initiatives. Jump straight to one of the sections below, or just scroll down to find out more.

Update (05/09/2022): We are running a SemEval competition and we release more sentiment dataset from African languages including NaijaSenti Dataset. Visit the AfriSenti SemEval page for more information : [AfriSenti-SemEval Task 12](#)

Update (02/10/2022): Send me email (shamsuddeen2004 at gmail.com) if you need more information about the dataset.

Download NaijaSenti Datasets

[Manually Annotated Twitter Sentiment Dataset](#)

[Manually Annotated Sentiment Lexicon](#)

[Semi-automatically Translated emotion lexicon](#)

[Semi-automatically Translated sentiment lexicon](#)

[Large Scale Unlabeled Twitter Sentiment Corpus](#)

[Stop-words for Hausa, Igbo, Pidgin and Yoruba](#)

IgboSynCorp Dataverse

[https://dataverse.harvard.edu/
dataverse/0000-0002-0068-6933](https://dataverse.harvard.edu/dataverse/0000-0002-0068-6933)

The screenshot displays the Harvard Dataverse interface for the IgboSynCorp Dataverse. At the top, the Harvard Dataverse logo is on the left, and navigation links for 'Add Data', 'Search', 'About', 'User Guide', 'Support', 'Sign Up', and 'Log In' are on the right. The main heading is 'IgboSynCorp Dataverse' with the affiliation '(University of Ibadan, Ibadan, Nigeria)'. Below this is a breadcrumb link 'Harvard Dataverse >' and 'Contact' and 'Share' buttons. A search bar contains the text 'Search this dataverse...' and an 'Advanced Search' button. On the left sidebar, there are filters for 'Dataverses (0)', 'Datasets (2)', and 'Files (126)'. Under 'Publication Year', '2022 (2)' is selected. Under 'Subject', 'Arts and Humanities (2)', 'Computer and Information Science (2)', and 'Other (2)' are listed. Under 'Author Name', 'Adiboshi, Anita (2)', 'Ejinwa, Samuel Obinna (2)', 'Ihunna Peter (2)', 'Nweya, Gerald Okey (2)', and 'Nwokwu, Daniel Success (2)' are listed. The main content area shows '1 to 2 of 2 Results' with a 'Sort' dropdown. The first result is 'Replication Data for Igbo Natural Language Processing Tasks I' dated Jul 8, 2022, by Nweya, Gerald Okey; Akinola Solomon Oluwole; Onwuegbuzia, Emeka Felix; Ejinwa, Samuel Obinna; Adiboshi, Anita; Nwokwu, Daniel Success; Ihunna Peter; Osuagwu, Amarachi Akudo, 2022. The second result is 'Replication Data for Igbo Natural Language Processing Tasks II' dated Jul 7, 2022, by the same authors. Both results include a brief description of the Igbo synchronised corpus.

HARVARD
Dataverse

Add Data ▾ Search ▾ About User Guide Support Sign Up Log In

[IgboSynCorp Dataverse](#)
(University of Ibadan, Ibadan, Nigeria)

[Harvard Dataverse >](#)

[✉ Contact](#) [🔄 Share](#)

Search this dataverse... [Advanced Search](#)

[Dataverses \(0\)](#)

[Datasets \(2\)](#)


[Files \(126\)](#)

Publication Year
2022 (2)


Subject
[Arts and Humanities \(2\)](#)
[Computer and Information Science \(2\)](#)
[Other \(2\)](#)

Author Name
[Adiboshi, Anita \(2\)](#)
[Ejinwa, Samuel Obinna \(2\)](#)
[Ihunna Peter \(2\)](#)
[Nweya, Gerald Okey \(2\)](#)
[Nwokwu, Daniel Success \(2\)](#)

1 to 2 of 2 Results [↑↓ Sort ▾](#)

Replication Data for Igbo Natural Language Processing Tasks I
Jul 8, 2022
 Nweya, Gerald Okey; Akinola Solomon Oluwole; Onwuegbuzia, Emeka Felix; Ejinwa, Samuel Obinna; Adiboshi, Anita; Nwokwu, Daniel Success; Ihunna Peter; Osuagwu, Amarachi Akudo, 2022, "Replication Data for Igbo Natural Language Processing Tasks I", <https://doi.org/10.7910/DVN/RXBNCZ>, Harvard Dataverse, V1, UNF:6:0RBYf/qoSeTZCEh9QK0xiQ== [fileUNF]

The Igbo synchronised corpus (IgboSynCorp) is an annotated corpus of spoken Igbo created by a team of linguists and NLP experts at the University of Ibadan and Afe Babalola University, Nigeria. The project was designed to create an open access labelled and unlabelled dataset for...

Replication Data for Igbo Natural Language Processing Tasks II
Jul 7, 2022
 Nweya, Gerald Okey; Akinola Solomon Oluwole; Onwuegbuzia, Emeka Felix; Ejinwa, Samuel Obinna; Adiboshi, Anita; Nwokwu, Daniel Success; Ihunna Peter; Osuagwu, Amarachi Akudo, 2022, "Replication Data for Igbo Natural Language Processing Tasks II", <https://doi.org/10.7910/DVN/YB9FWK>, Harvard Dataverse, V1

The Igbo synchronised corpus (IgboSynCorp) is an annotated corpus of spoken Igbo created by a team of linguists and NLP experts at the University of Ibadan and Afe Babalola University, Nigeria. The project was designed to create an open access labelled and unlabelled dataset for...

IgboSentilex

Development of a General Purpose Sentiment Lexicon for Igbo Language

Emeka Ogbuju

Department of Computer Science, Federal
University Lokoja, Nigeria
emeka.ogbuju@fulokoja.edu.ng

Moses Onyesolu

Department of Computer Science, Nnamdi
Azikiwe University, Awka, Nigeria
mo.onyesolu@unizik.edu.ng

Abstract

There are publicly available general purpose sentiment lexicons in some high resource languages but very few exist in the low resource languages. This makes it difficult to directly perform sentiment analysis tasks in such languages. The objective of this work is to create a general purpose sentiment lexicon for the Igbo language that can determine the sentiment of documents written in the Igbo language without having to translate it to the English language. The material used was an automatically translated Liu's lexicon and manual addition of Igbo native words. The result of this work is a general purpose lexicon – *IgboSentilex*. The performance was tested on the BBC Igbo news channel. It returned an average polarity agreement of 95.75% with other general purpose sentiment lexicons.

1. Introduction

Sentiment analysis or opinion mining is a natural language processing task that deals with the determination of positive, negative or neutral polarities of texts such as news articles, blogs, reviews or

Yoruba NLP Resources

awesome-yoruba-nlp



A curated list of resources dedicated to Natural Language Processing in the Yoruba Language.

Maintainers - [Olamilekan Wahab](#)

Please read the [contribution guidelines](#) before contributing.

Please feel free to create [pull requests](#).

Contents

- [Papers](#)
- [Slides](#)
- [Datasets](#)

Papers

- [Natural Language Processing of English To Yoruba](#)
- [A computerized identification system for verb sorting and arrangement in a natural language : Case Study of the Nigerian Yoruba Language](#)
- [Morphological Analysis of Standard Yorùbá Nouns](#)
- [Part-of-Speech tagging of Yoruba Standard, Language of Niger-Congo family](#)
- [Rule Based Parts of Speech Tagging of Yorùbá Simple Sentences](#)

Postgraduate Programmes in NLP

IFRA-Nigeria

Institut Français de Recherche en Afrique
French Institute for Research in Africa
Institute of African Studies, University of Ibadan

HOME

ABOUT US

NEWS

RESEARCH PROJECTS

DIGITAL HUMANITIES

EVENTS

PUBLICATI

[Home](#) / [Events](#) / [Training](#) / [Masterclasses](#) / [ADEFSA Natural Language Processing \(NLP\) Bootcamp](#)

ADEFSA Natural Language Processing (NLP) Bootcamp

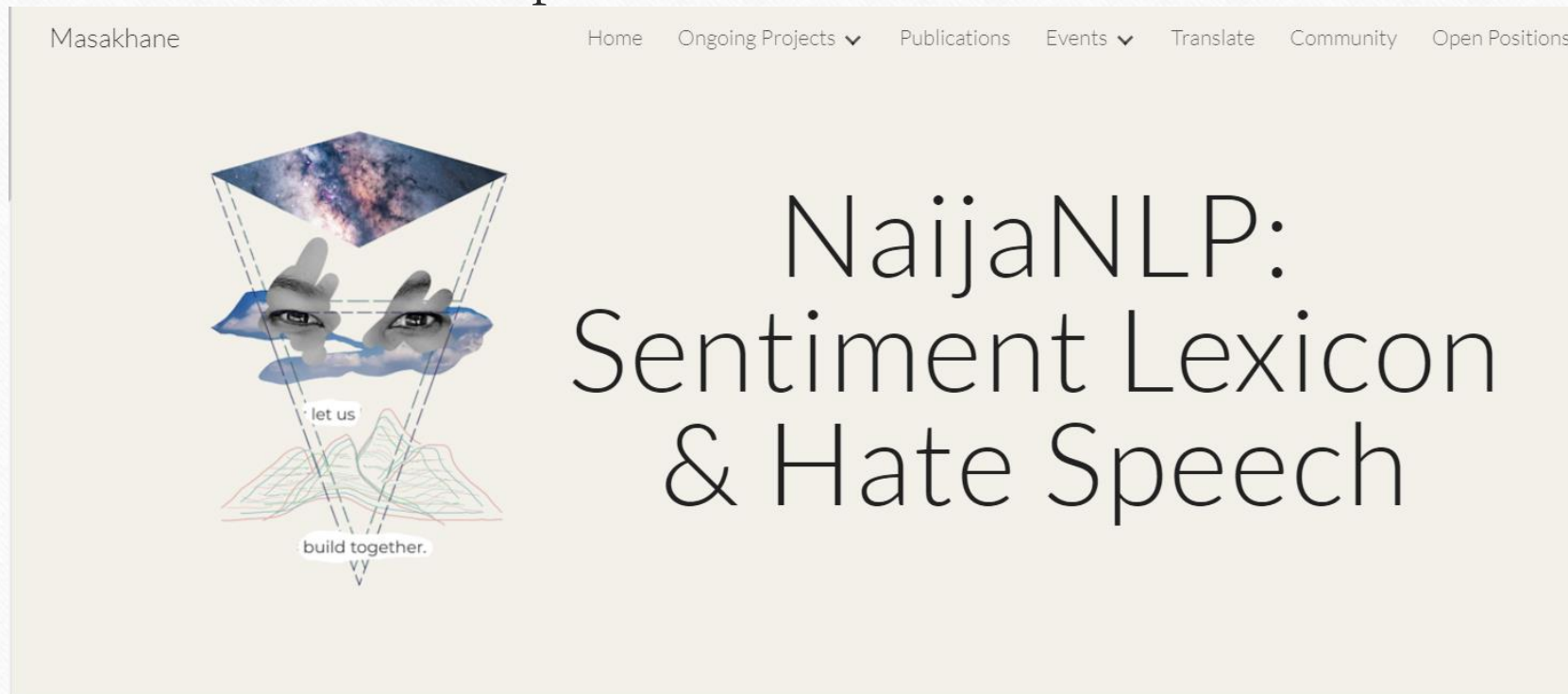


The **ADEFSA Natural Language Processing (NLP) Bootcamp** was run by IFRA from the 8th to the 11th of November, as part of a programme to support the development of French higher education in Africa (ADESFA) financed by the French Ministry of Europe and Foreign Affairs (MEAE). ADEFSA has supported the University Paris Nanterre to help the University of Ibadan to set up a Masters Degree in Natural Language Processing.

University of Ibadan Master's degree. This degree will prepare students from the humanities as well as the exact sciences to be operational in the automatic information processing and linguistic engineering sector. The skills acquired belong to **linguistics, statistics and computer sciences**. The target audience is students who are not yet specialized in ICST, with diverse academic backgrounds (humanities, linguistics, statistics and computer science, etc.) and who are destined for various careers (communication, business, police and security, higher education and research, etc.).

Masakhane

<https://www.masakhane.io/>



Pos-Tagging/NER Datasets

[MasakhaPOS: Part-of-Speech Tagging Dataset for 20 African Languages](#)

MasakhaPOS is the largest human-annotated part of speech tagging dataset for 20 African languages. Each language has between 1200 – 1500 sentences for training and/or evaluation. The languages covered span across West, Central, East and Southern Africa, and include Bambara, Ghomala, Ewe, Fon, Hausa, Igbo, Kinyarwanda, Luganda, Dholuo, Mossi, Chichewa, Nigerian-Pidgin, chiShona, Setswana, Swahili, Twi, Wolof, isiXhosa, Yorùbá, and isiZulu.

Named Entity Recognition and Parts of Speech datasets for African languages

CONTACT: DAVID IFEOLUWA ADELANI, D.ADELANI@UCL.AC.UK

[MasakhaNER 2.0: Named Entity Recognition datasets for 20 African languages](#)

MasakhaNER 2.0 is the largest human-annotated named entity recognition dataset for 20 African languages. Each language has between 4,800 – 11,000 parallel sentences for training and/or evaluation. The languages covered span across West, Central, East and Southern Africa, and include Bambara, Ghomala, Ewe, Fon, Hausa, Igbo, Kinyarwanda, Luganda, Dholuo, Mossi, Chichewa, Nigerian-Pidgin, chiShona, Setswana, Swahili, Twi, Wolof, isiXhosa, Yorùbá, and isiZulu. More information about the data can be found in their EMNLP paper [here](#).

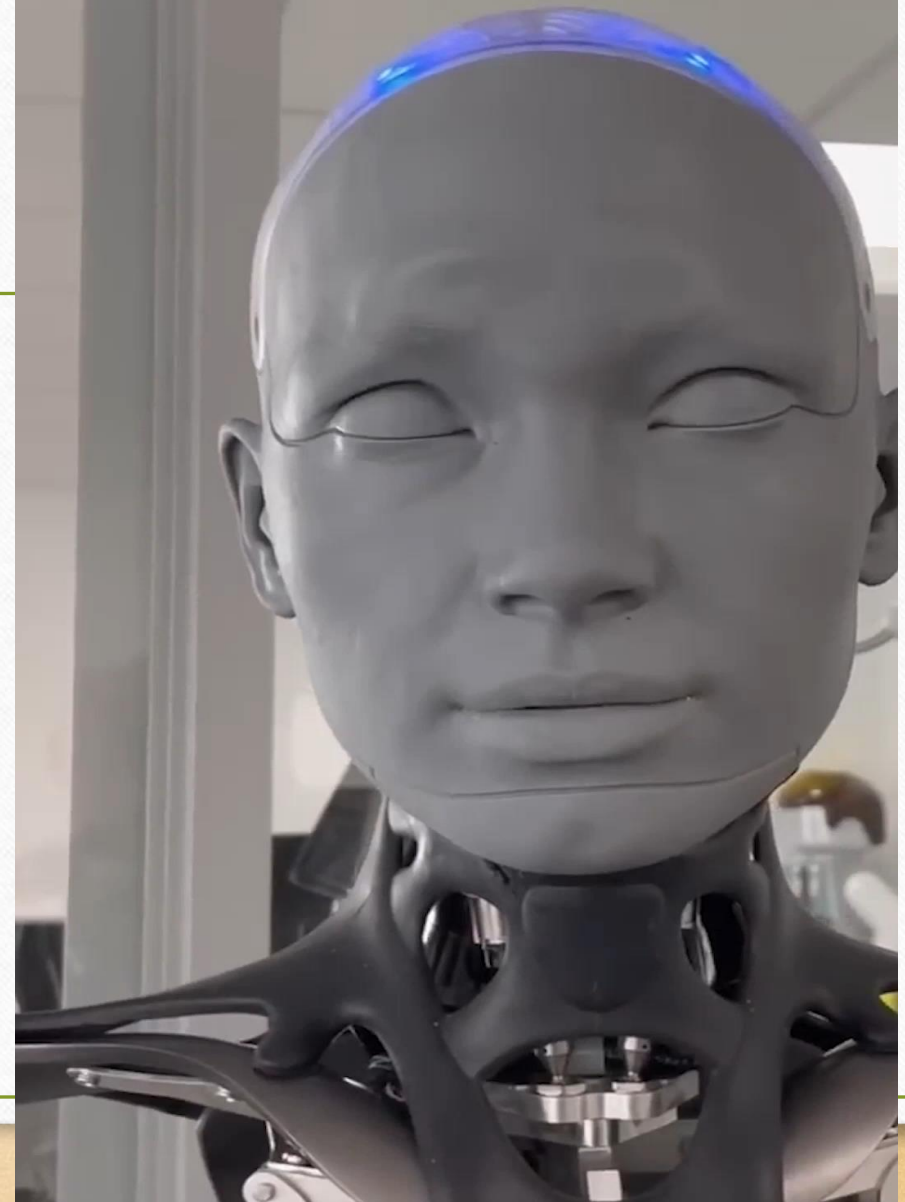
Indigenous or Native Robots

“The 4th Industrial revolution in Africa cannot take place in English. It is imperative that NLP models be developed for the African continent”

- Masakhane

Ameca speaks German,
English, French, Japanese,
& Chinese using OpenAI
ChatGPT

Source: Engineered Arts



Nigeria's 1st Humanoid

Omeife speaks local languages



UNICCON GROUP UNVEILS OMEIFE HUMANOID ROBOT

South Africa's 1st Humanoid

English?



The Setup – 4DBI

1. **Define requirements specific to a domain** - communication, education, entertainment, health, etc.
2. **Design a Robot for that domain (not necessarily a humanoid)**
3. **Develop a Speech Recognition System for it – focus on a language!**
4. **Develop a models for Natural Language Understanding** - common phrases, grammar rules, and semantic relationships to extract intents, entities, and context from your language inputs.
5. **Build Language Generation System – that can implement** text-to-speech synthesis to convert textual Igbo representations into spoken speech.
6. **Integrate a Speech Interaction System – for** dialogue management mechanisms for turn-taking, context retention, and error handling for a cohesive speech interaction system.

Key Take-Aways

- It is our responsibility to train our own robots to understand our language
- We must curate local datasets and make them available/open for research
- We must fund the curation also
- Universities should float Masters Degree Programme in NLP
- Students in MSc Programmes in Computer Science should embark on NLP projects
- Tech companies should collaborate with research institutes for advanced language models

References

- Naijasenti: A Nigerian twitter sentiment corpus for multilingual sentiment analysis, SH Muhammad, DI Adelani, S Ruder, IS Ahmad, I Abdulmumin, BS Bello, M Choudhury... arXiv preprint arXiv:2201.08277, 2022•arxiv.org
- Nweya, Gerald Okey; Akinola Solomon Oluwole; Onwuegbuzia, Emeka Felix; Ejinwa, Samuel Obinna; Adiboshi, Anita; Nwokwu, Daniel Success; Ihunna Peter; Osuagwu, Amarachi Akudo, 2022, "Replication Data for Igbo Natural Language Processing Tasks I", <https://doi.org/10.7910/DVN/RXBNCZ>, Harvard Dataverse, V1, UNF:6:0RBYf/qoSeTZCEh9QKOxiQ== [fileUNF]
- <https://www.ifra-nigeria.org/events/training/masterclasses/444-adeftsa-natural-language-processing-nlp-bootcamp>
- <https://github.com/masakhane-io/masakhane-ner/tree/main/MasakhaNER2.0/data>
- <https://github.com/masakhane-io/masakhane-pos>
- <https://github.com/Olamyy/awesome-yoruba-nlp>
- <https://nypost.com/2023/04/10/worlds-most-advanced-robot-speaks-several-languages-in-creepy-video/>
- [https://africachinapresscentre.org/2022/11/13/nigerian-tech-company-makes-first-african-humanoid-robot-that-speaks-over-6-languages-displays-african-ethics/-](https://africachinapresscentre.org/2022/11/13/nigerian-tech-company-makes-first-african-humanoid-robot-that-speaks-over-6-languages-displays-african-ethics/)
- <https://www.bbc.co.uk/sounds/play/p0d5h6h9>

Q/A



thank

you