NaijaSenti

Hausa-NLP Research Group

Lacuna Fund
Our voice on data

U.PORTO

NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis

Shamsuddeen Muhammad(@shmuhammadd)
University of Porto, Bayero University, Kano - Nigeria, MasakhaneNLP, HausaNLP

NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis

Shamsuddeen Hassan Muhammad[1,2*+], **David Ifeoluwa Adelani**[3*], **Sebastian Ruder**[4]
Ibrahim Sa'id Ahmad[5+], **Idris Abdulmumin**[6*+], **Bello Shehu Bello**[5+], **Monojit Choudhury**[7]
Chris Chinenye Emezue[8*], **Saheed Salahudeen Abdullahi**[10+]
, Anuoluwapo Aremu[11*], **Alipio Jeorge**[1,2] **Pavel Brazdil**[1]

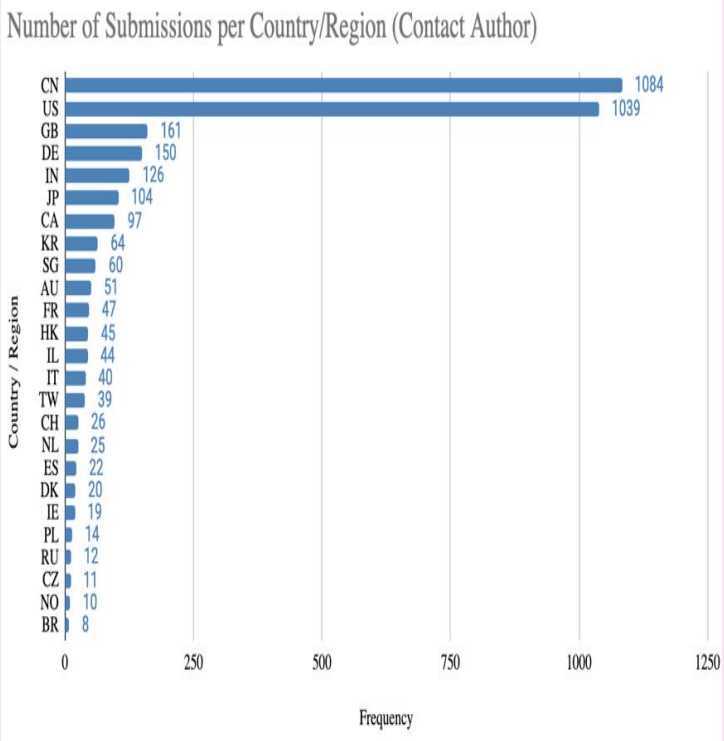
[1] LIAAD - INESC TEC, [2]Faculty of Sciences-University of Porto, Portugal,
[3]Spoken Language Systems Group (LSV), Saarland University, Germany, [4]Google Research
[5]Faculty of Computer Science and Information Technology, Bayero University, Kano, Nigeria
[6]Department of Computer Science, Ahmadu Bello University, Zaria, Nigeria, [7]Microsoft Research India,
[8]Technical University of Munich, Germany, [9]Clear Global, [10]Kaduna state University
[*]MasakhaneNLP ,[+]HausaNLP

Table of contents

Motivation

30% of all living languages today are African languages

(Ethnologue)

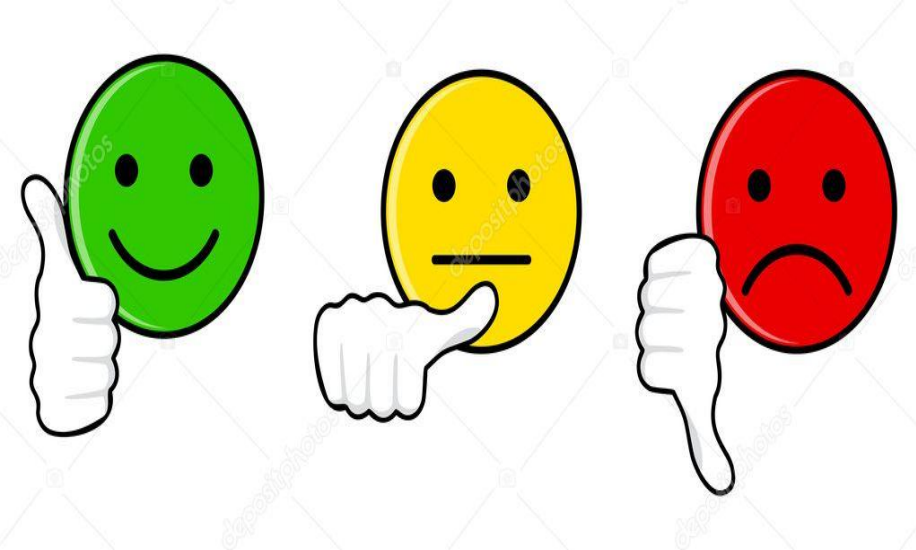


https://acl2020.org/blog/general-conference-statistics/

02/04

Sentiment Analysis for low-resource

Sentiment Analysis for Low-resource



SA involves finding opinion or sentiment in a text !

- Considerable increase of interest
- English centric
- Low-resource:
  - Lack of available annotated data
  - Translation
    - Focus > English
    - English > Focus

Translation Alter Sentiment

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

SEMANTIC ENRICHMENT OF NIGERIAN PIDGIN ENGLISH FOR CONTEXTUAL SENTIMENT CLASSIFICATION.

Wuraola Fisayo Oyewusi
Data Science Nigeria
Lagos, Nigeria
wuraola@datasciencenigeria.ai

Olubayo Adekanmbi
Data Science Nigeria
Lagos Nigeria
olubayo@datasciencenigeria.ai

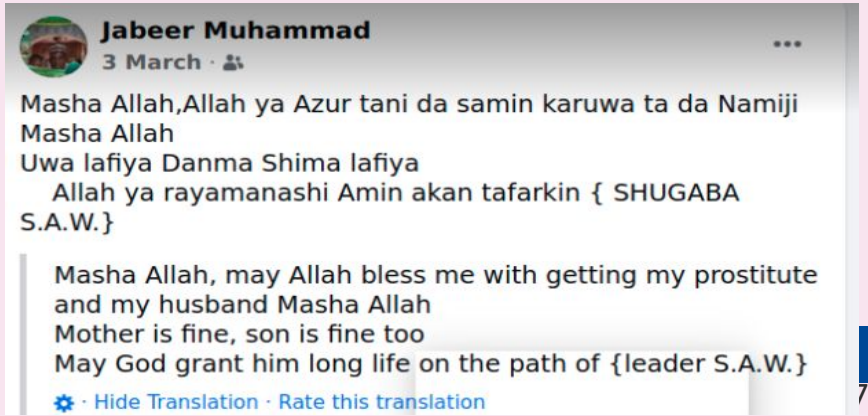
Olalekan Akinsande
Data Science Nigeria Lagos,Nigeria
olalekan@datasciencenigeria.ai

How Translation Alters Sentiment

Saif M. Mohammad
National Research Council Canada
SAIF.MOHAMMAD@NRC-CNRC.GC.CA

Mohammad Salameh
University of Alberta
MSALAMEH@UALBERTA.CA

Svetlana Kiritchenko
National Research Council Canada
SVETLANA.KIRITCHENKO@NRC-CNRC.GC.CA



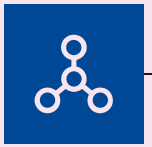
Previously Annotated Datasets

Dataset	Language	Open-source	Annotated/translated	Code-mixed	Source
Abubakar et al. (2021)	Hausa	✗	annotated	✓	Twitter
Ogbuju and Onyesolu (2019)	Igbo	✗	translated	✗	General
Umoh et al. (2020)	Igbo	✗	annotated	✗	General
Oyewusi et al. (2020)*	Pidgin	✗	annotated/translated	✓	Twitter
Orimaye et al. (2012)	Yorùbá	✓	annotated	✓	Youtube
Iyanda and Abegunde (2019)	Yorùbá	✗	annotated	✗	General
Ours	Hausa, Igbo, Yorùbá, Pidgin	✓	annotated	✓	Twitter

Table 1: Summary of datasets used in six existing datasets on sentiment analysis in four major Nigerian languages in comparison to ours. *: The provided URL is no longer accessible.

NaijaSenti Contributions

1st



Labelled Dataset

Labelled tweets in four Nigerian languages (Hausa, Yoruba, Igbo, Pidgin).

2nd

Unlabbed Dataset

Released the largest Twitter corpus in four languages

3rd

Sentiment Lexicon

Released annotated sentiment lexicon and stopword in Hausa, Yoruba and Igbo

4th

Trained Models

Train various PLM and Released the code

03/04

Data collection and Annotation

Dataset collection

Twitter Academic API

- Support 70 languages (only one African lang: Amharic)
- None among the Nigerian languages is supported

Stopwords, emoji, sentiment lexicon

- Stopword + Emoji
- Translated sent.lexicon
- Hashtag e.g #Yorubaday
- Handles e.g @bbchausa

Language Detection

- Stopword overlap
- “*nke*”, produce tweet in Hausa *amin ya rabbi godiya nke*”
- longitude, latitude, and radius parameters (25 miles)

Annotation Process

1

Five-way

Pos, neg, neu, mix, ind

2

Annotation tool

Light-tag

3

Annotators

3-native speakers

4

Training

- Iteratively
- CS and Linguist

5

Tweet Annotation

- Batches (1000/batch)
- Approx **30,000** tweets

6

Adjudication

Disagreement cases

Inter-annotators progress over Thirty batches

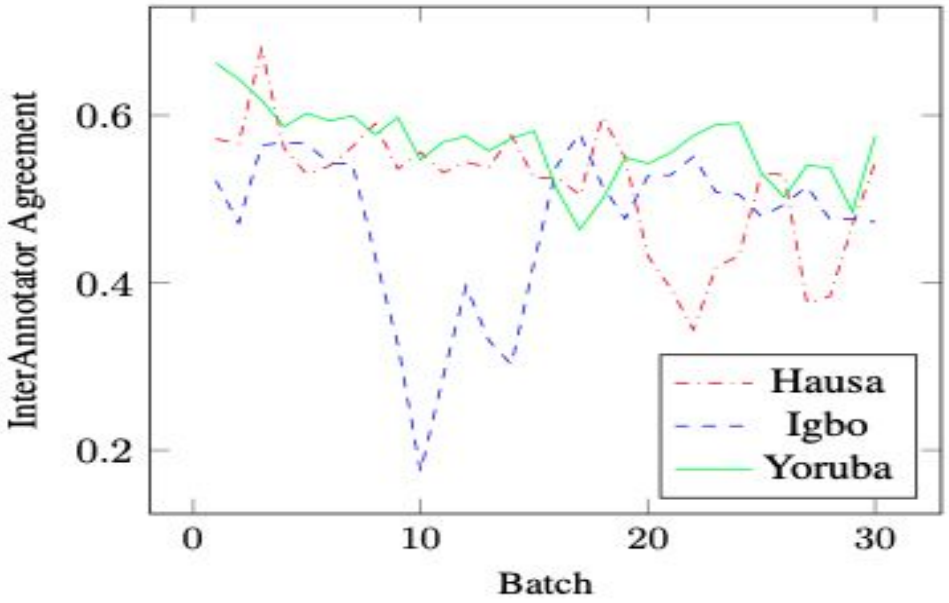
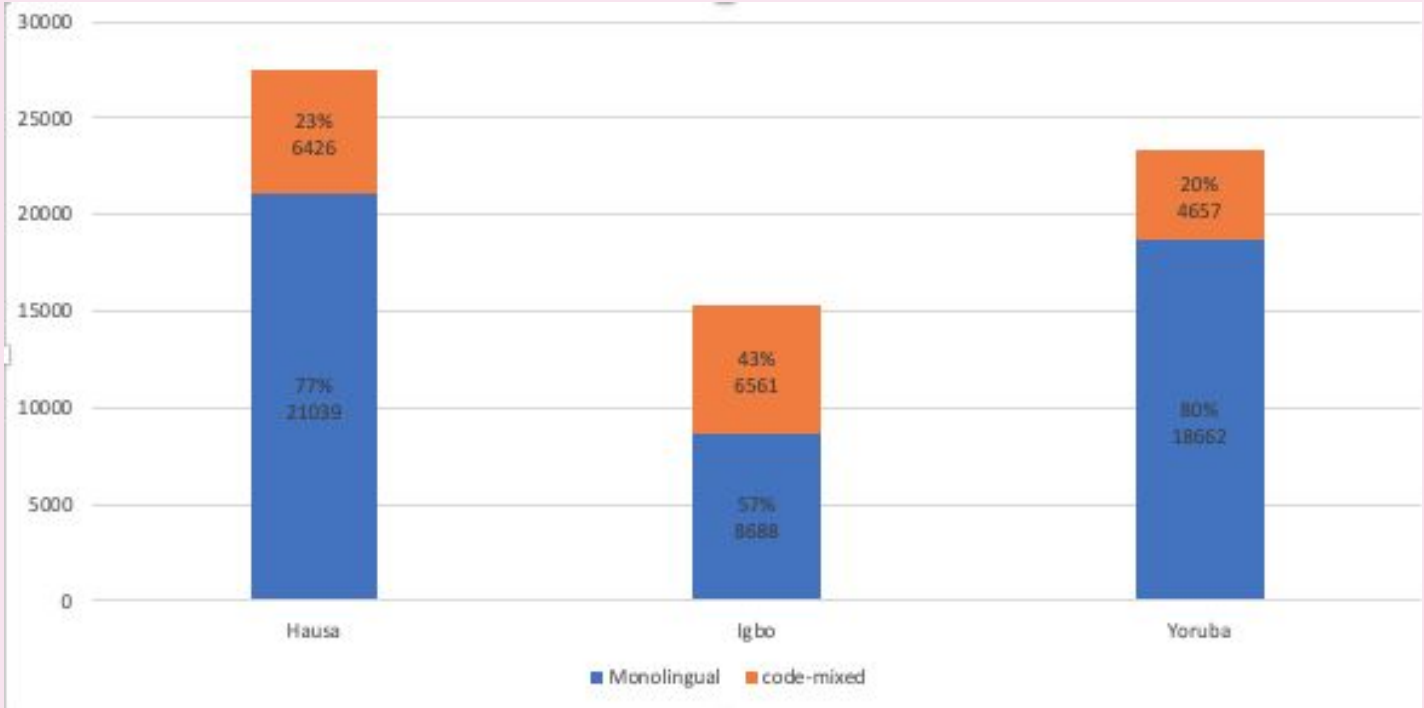


Figure 1: Inter-annotators progress over thirty batches—one-thousand tweets per batch.

Monolingual vs Code-mix



Diacritics challenge

E.g.,Yoruba

- *“Awon omo fo abo”* does not have a meaning without diacritics,
- annotators classify it as indeterminate

Tonality

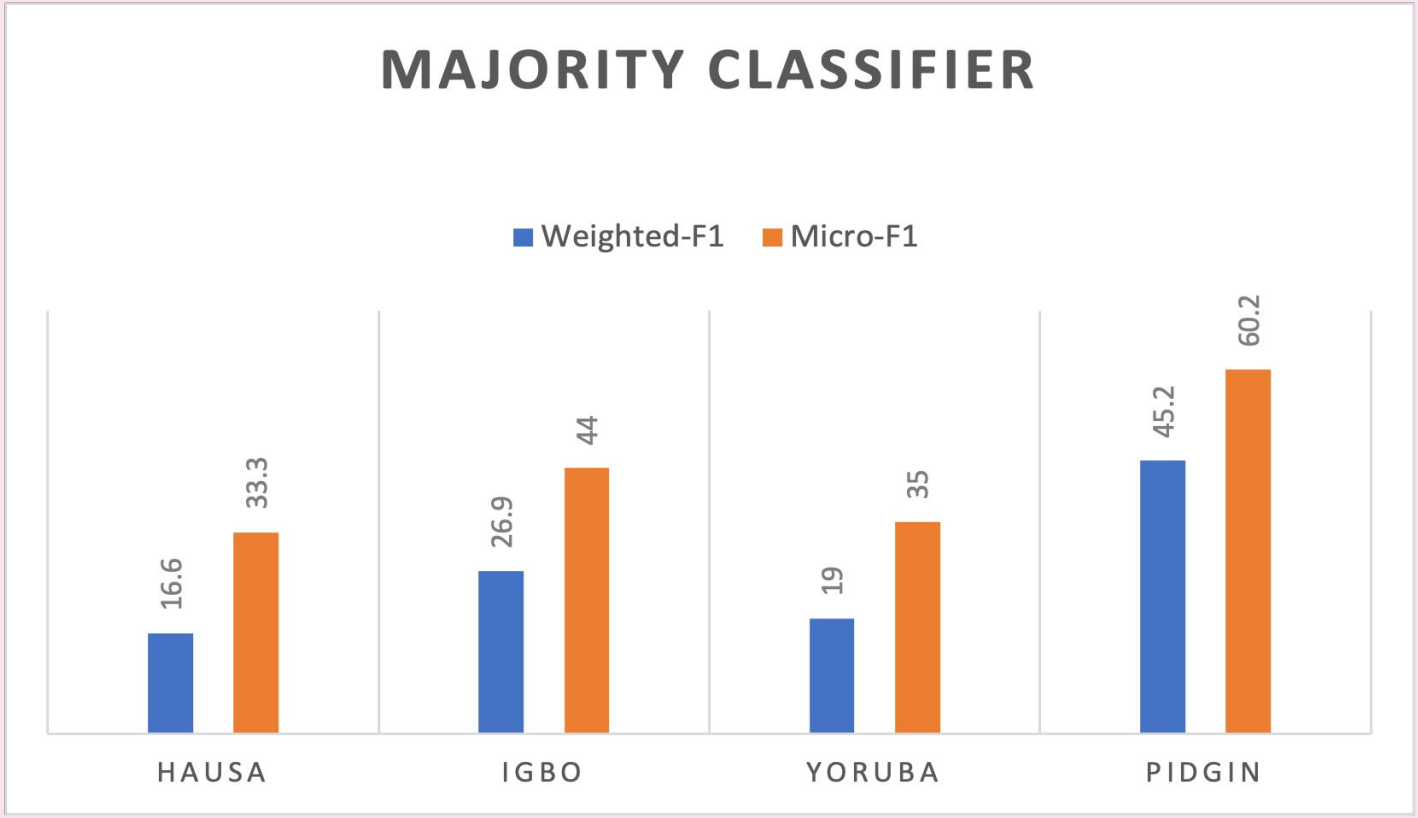
- Tone in Yorùbá and Igbo helps to give meaning to words in context,
- The same sentence but different tones may have opposite sentiment:
  - Àwon omó fo abó (The children washed the dishes) has a **positive** meaning,
  - Àwon omó fó abó ..(The children broke the dishes) has a **negative** meaning

Tonality

- Similarly, tonality is heavily used in Igbo.
  - ò nwèkwàrà mgbe i naenwe sense ? – Will you ever be able to talk sensibly? – You’re a fool.
  - • ò nwèkwàrà mgbe i naenwe sense – Sometimes you act with great maturity – I’m impressed.

04

Benchmark Experiments


MAJORITY CLASSIFIER
Weighted-F1
Micro-F1
16.6
33.3
26.9
44
19
35
45.2
60.2
HAUSA
IGBO
YORUBA
PIDGIN

Multilingual PLMs

AfriBERTa winner

Model	NG lang. supported	PLM size	hau	ibo	pcm	yor	Avg
Multilingual PLMs							
AfriBERTa-large	hau, ibo, pcm, yor	126M	$81.0^{*}_{\pm 0.2}$	$\mathbf{81.2^{*}_{\pm 0.5}}$	$75.0_{\pm 0.6}$	$80.2^{*}_{\pm 0.6}$	$79.3^{*}_{\pm 0.3}$
mBERT-base	yor	172M	$77.8_{\pm 0.5}$	$79.8_{\pm 0.5}$	$72.4_{\pm 1.5}$	$77.6_{\pm 0.9}$	$76.9_{\pm 0.3}$
XLM-R-base	hau	270M	$78.4_{\pm 1.0}$	$79.9_{\pm 0.7}$	$76.3_{\pm 0.6}$	$76.9_{\pm 0.4}$	$77.9_{\pm 0.2}$
mDeBERTaV3-base	None	276M	$79.3_{\pm 0.}$	$80.7_{\pm 0.2}$	$77.6^{*}_{\pm 0.8}$	$78.4_{\pm 0.5}$	$79.0_{\pm 0.3}$
RemBERT	hau, ibo, yor	559M	$79.0_{\pm 0}$	$79.9_{\pm 0.4}$	$\mathbf{78.4^{*}_{\pm 1.4}}$	$78.0_{\pm 0.6}$	$78.8_{\pm 0.6}$

Better than others

Language adaptive fine-tuning (LAFT)

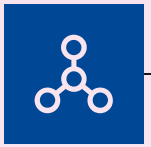
AfriBERTa
Competitive with
LAFT

Model	NG lang. supported	PLM size	hau	ibo	pcm	yor	Avg
Multilingual PLMs+LAFT							
mBERT+LAFT (General)	hau / ibo / pcm / yor	172M	$80.8_{\pm 0.3}$	$80.4_{\pm 0.4}$	$74.2_{\pm 0.5}$	$80.8_{\pm 0.5}$	$79.1_{\pm 0.3}$
mBERT+LAFT (Tweet)	hau / ibo / pcm / yor	172M	$79.3_{\pm 0.6}$	$77.7_{\pm 0.6}$	$74.0_{\pm 0.7}$	$76.8_{\pm 0.3}$	$77.0_{\pm 0.3}$
XLM-R-base+LAFT (General)	hau / ibo / pcm / yor	270M	$\mathbf{81.5^{*}_{\pm 0.7}}$	$80.8^{*}_{\pm 0.8}$	$74.7_{\pm 1.5}$	$\mathbf{80.9^{*}_{\pm 0.4}}$	$\mathbf{79.5^{*}_{\pm 0.3}}$
XLM-R-base+LAFT (Tweet)	hau / ibo / pcm / yor	270M	$79.5_{\pm 0.9}$	$77.0_{\pm 0.5}$	$74.8_{\pm 0.7}$	$76.2_{\pm 0.4}$	$76.9_{\pm 0.2}$

- CC100 > 318MB general purpose
- Hausa tweets =32MB

Sentiment classification model

1st step



Multilingual PLMs

(mBert, XLM-R, AfriBERTa, mDeBER-TaV3, Remberta)

2nd step

Language adaptive fine-tuning (LAFT)

3rd step

Multi-task sentiment classification

4th step

Zero-shot Cross-Lingual Transfer

Multi-task Sentiment Classification

Model	NG lang. supported	PLM size	hau	ibo	pcm	yor	Avg
Multi-task Multilingual PLMs							
AfriBERTa-large	hau, ibo, pcm, yor	126M	$81.2^{*}_{\pm 0.2}$	$80.8^{*}_{\pm 0.2}$	$74.5_{\pm 0.6}$	$80.4^{*}_{\pm 0.7}$	$79.3_{\pm 0.3}$
mDeBERTaV3-base	None	276M	$79.0_{\pm 0.2}$	$79.3_{\pm 0.5}$	$76.8_{\pm 0.6}$	$78.7_{\pm 0.4}$	$78.4_{\pm 0.3}$

AfriBERTa winner

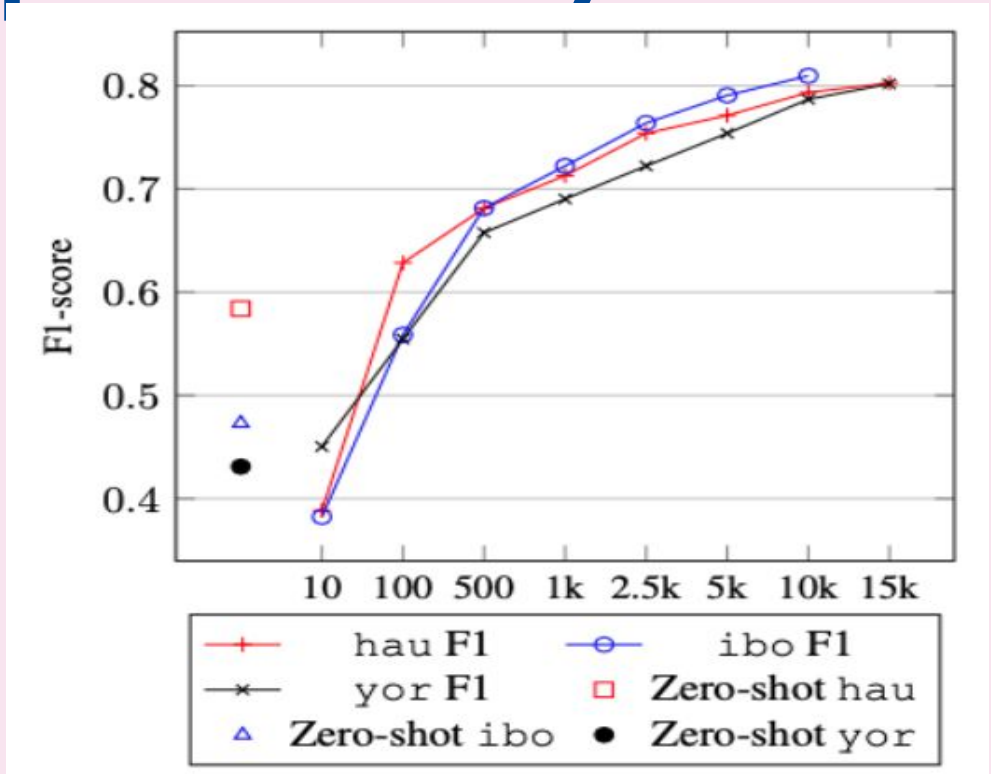
mDeBERTaV3-base
sight drop (−0.6%)

Zero-shot Cross-Lingual Transfer

Model	hau	ibo	pcm	yor	Avg
AfriBERTa-large	**58.4**	**47.7**	62.0	**43.1**	**52.8**
mBERT-base	31.0	37.0	57.4	39.5	41.2
XLM-R-base	38.4	37.8	62.4	26.7	41.3
mDeBERTaV3-base	50.1	47.2	64.9	36.4	49.7
RemBERT	54.0	45.4	**66.2**	30.2	49.0

Table 8: Transfer Learning experiments. PLMs are trained on English SemEval 2017 and evaluated on NG languages in a zero-shot setting

Sample Efficiency in Transfer



Conclusions and Future Work

- We present NaijaSenti
- Sentiment Lexicon
- Benchmark experiments
  - AfriBERTa
- NaijaSenti has the potential to spark interest in sentiment analysis and other downstream NLP tasks in the languages involved.
- Extend the corpus : AfriSenti

Thanks!

Paper: https://arxiv.org/abs/2103.11811

Code: https://github.com/hausanlp/NaijaSenti