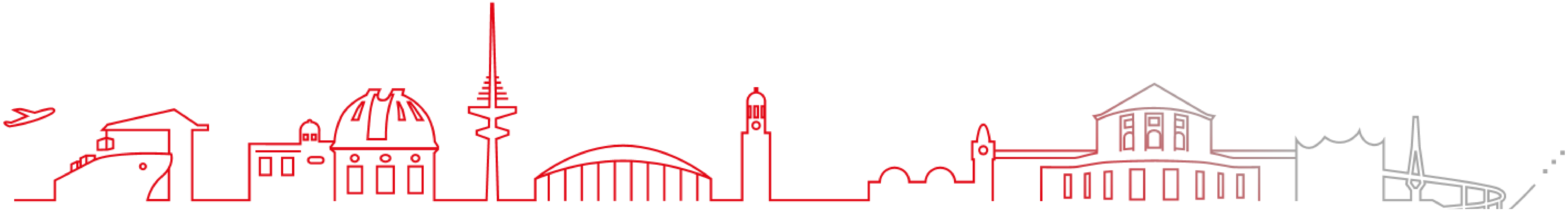


Modeling Language and Low-Resource Humanities Data

Lessons from African Low-Resource NLP for Humanities Research

Hub of Computing and Data Science
University of Hamburg

Seid Muhie Yimam

24.09.2025

From your photo, it looks like yo
or bus) because of the motion b
sandy ground and wire fencing –
Berlin.

So my best guess: you're somew
route.

Do you want me to try narrowing

ChatGPT – P

From your
bus) becau
ground an
So my bes
Do you wa

(likely a train or
orest with sandy
ound Berlin.
bly on a train route.

Hausa

Mun sami karuwa
(we got new baby)

English

We got prostitute

Agenda

- Introduction to HCDS
- Low-Resource Research: Datasets, Tools, and Benchmarks
- Modelling Language and Low-Resource
- Interdisciplinary Collaboration and Co-Creation - Tools
- Challenges and Future Directions

- This presentation includes contributions from my **EthioNLP** and **HausaNLP** team members, as part of our collaborative work on low-resource languages. Some slides have also been adapted from our joint materials.

Introduction - HCDS

Motivation: DFG-Position paper 2020
Digitale Transformation in Research

Levels of digital transformation in science and the humanities:

- **transformative change:** transfer of analog information and practices in the digital space
- **enabling change:** use of data-intensive technologies to address research questions
- **substituting change:** digitally support substantial amounts of the research process; redefine the process

Main road to happiness: scaling up!

Clearly, this should happen. But – how?

https://zenodo.org/record/4191345#.X70Wnz-g9aQ%23.X70Wnz-g9aQ

Hub of Computing and Data Science

HCDS
HUB OF COMPUTING
& DATA SCIENCE

central unit of the University of Hamburg

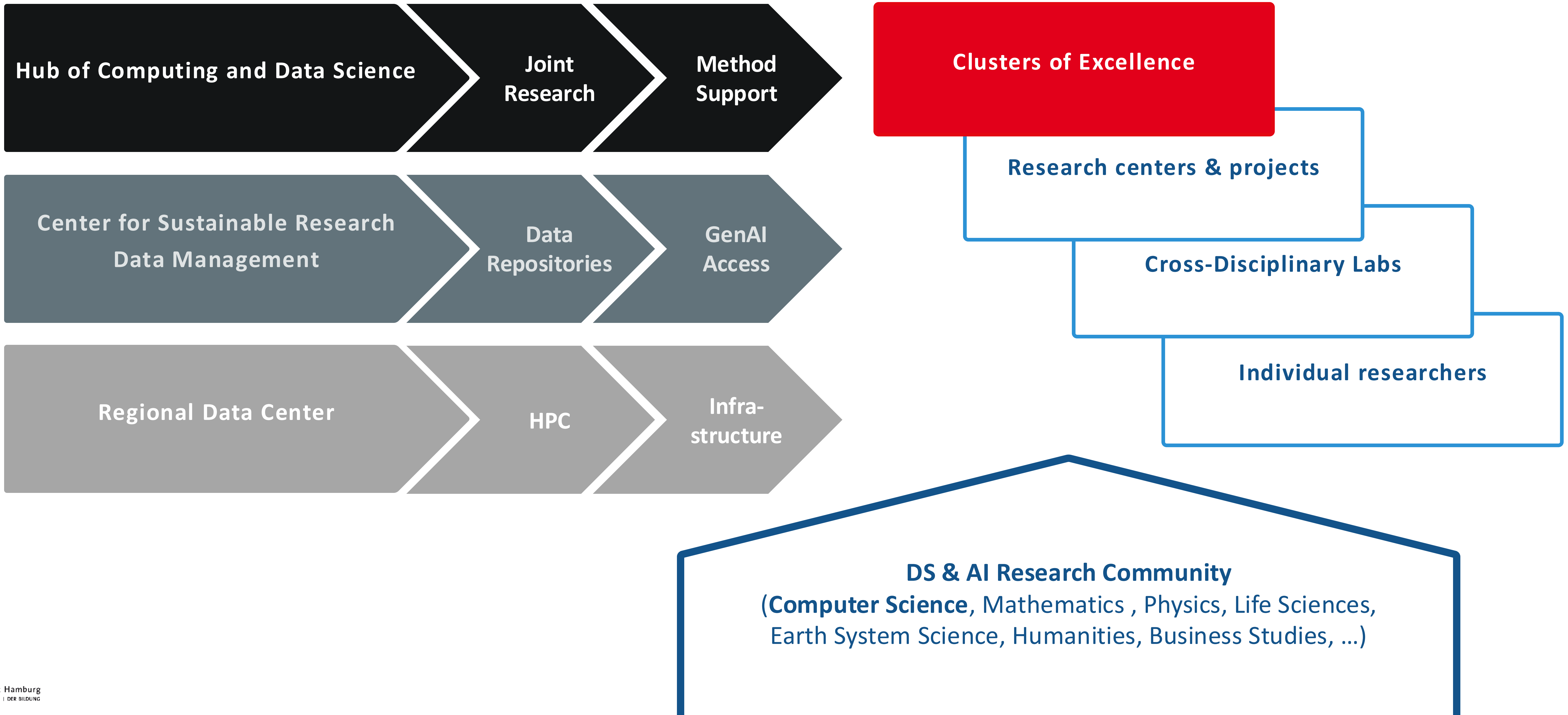
supports interdisciplinary research and application of innovative digital methods

coordinates and supports the implementation of the digital strategy in research at the Universität Hamburg

fuels adoption, use, and research of digital methods with Methodology Competence Centre

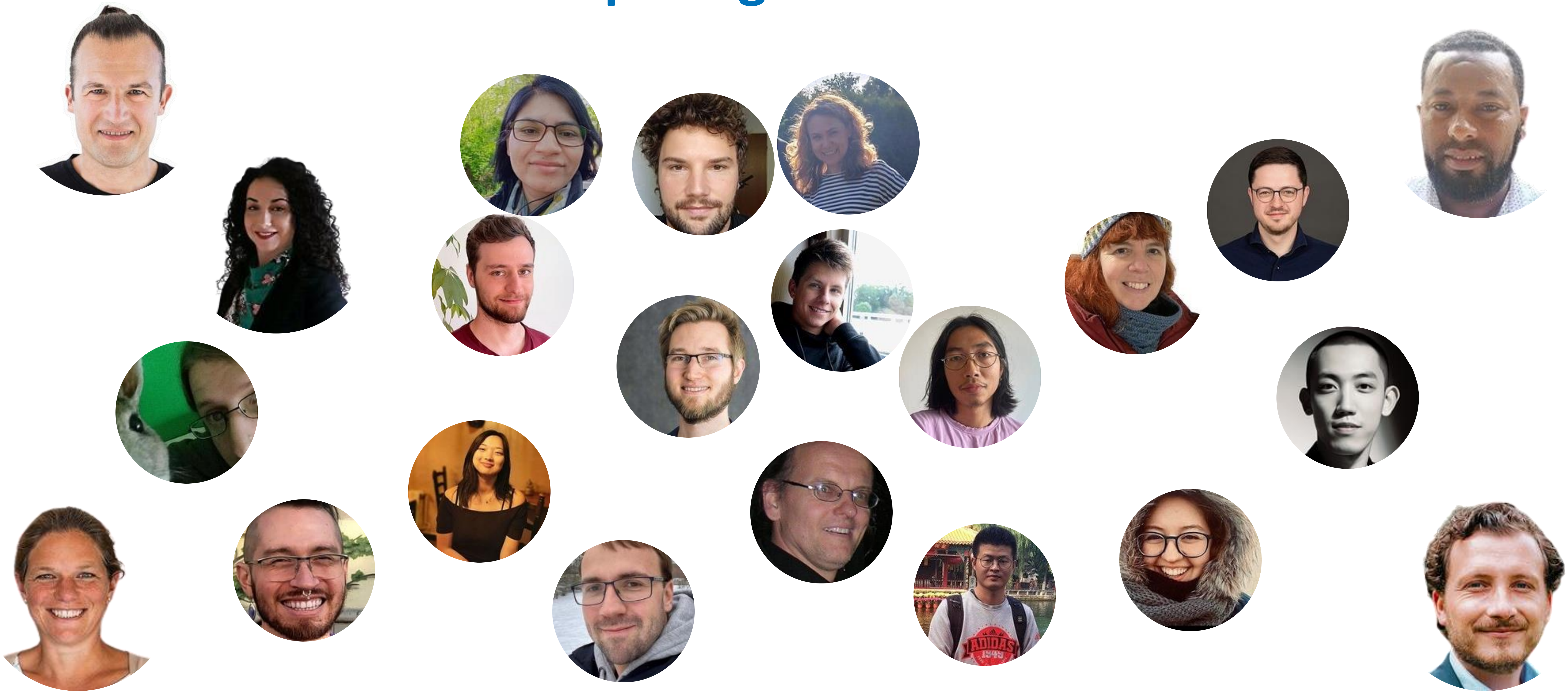
offer a forum for the exchange of information and collaboration at the interface between methodological sciences and applied sciences in the Cross-Disciplinary Labs

Central Units for Digital Matters in Research at University of Hamburg



Hub of Computing and Data Science

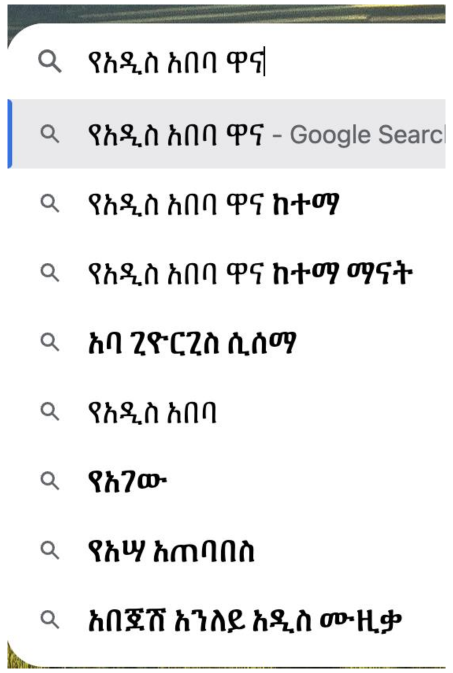
HCDS
HUB OF COMPUTING & DATA SCIENCE



Low-Resource Research: Datasets, Tools, and Benchmarks

Why research on other languages?

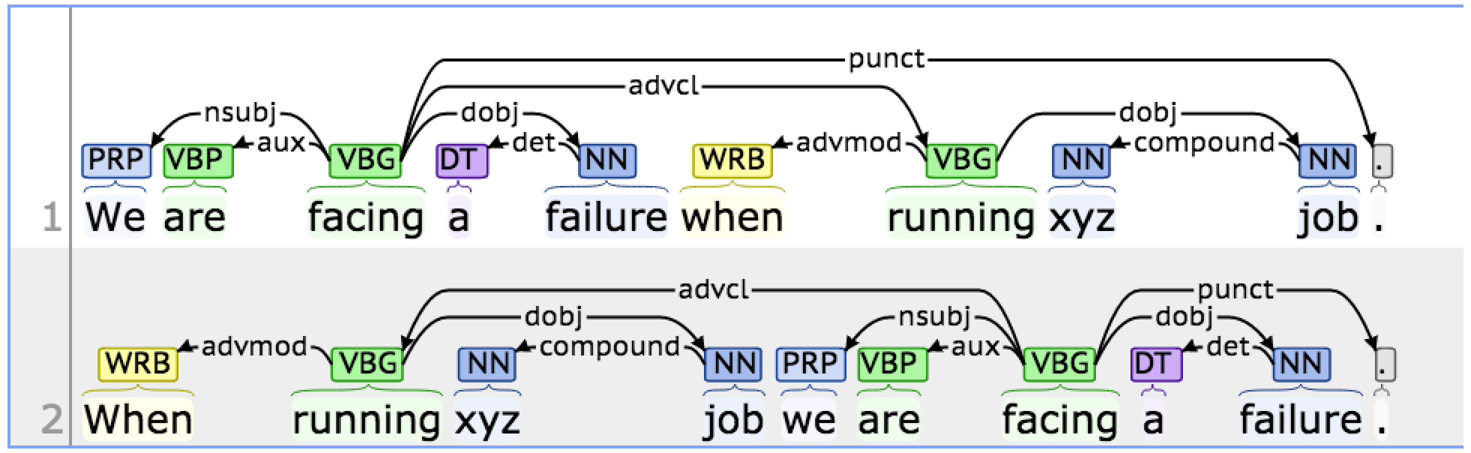
Information access in under-resourced languages:
e.g. how does one ….?



Enhance human communication e.g. machine translation

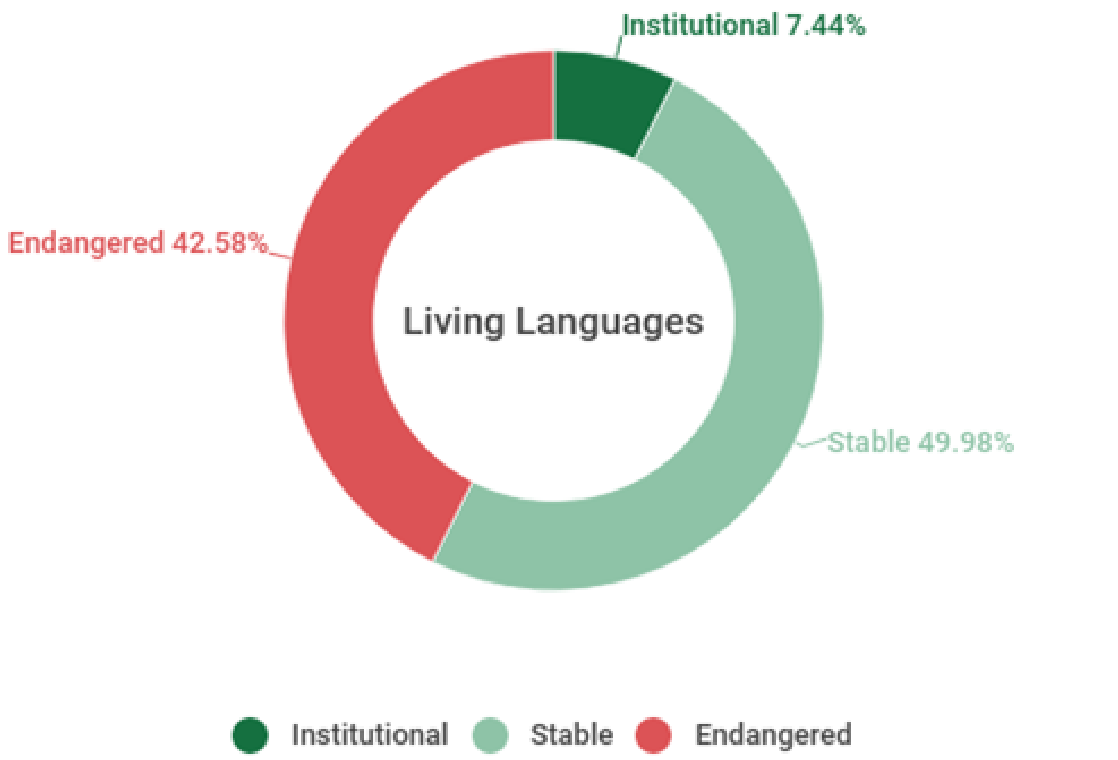
- **Enhance voice interaction: human-machine**
- **E.g Text-to-Speech, Speech-to-text, speech translation, dialog systems**
- **A lot of potential in education, tourism, business, humanitarian responses …**

Understanding linguistic structures of various languages and saving languages from dying

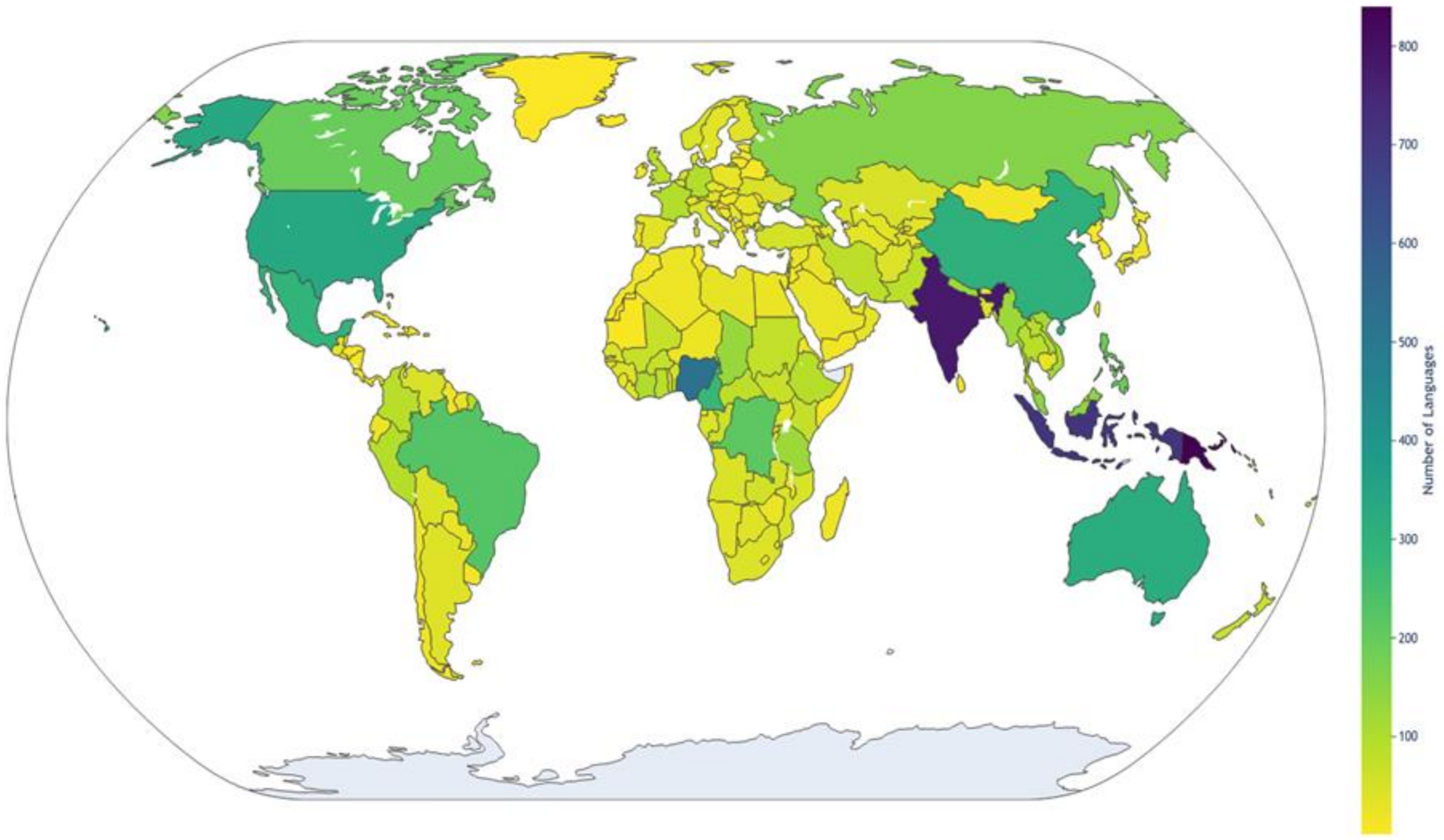


Linguistic Diversity

- **7,151** known languages (Ethnologue; 2022).
- ≈**400** have more than **1M speakers.**
- ≈**1,200 languages** have more than **100k.**



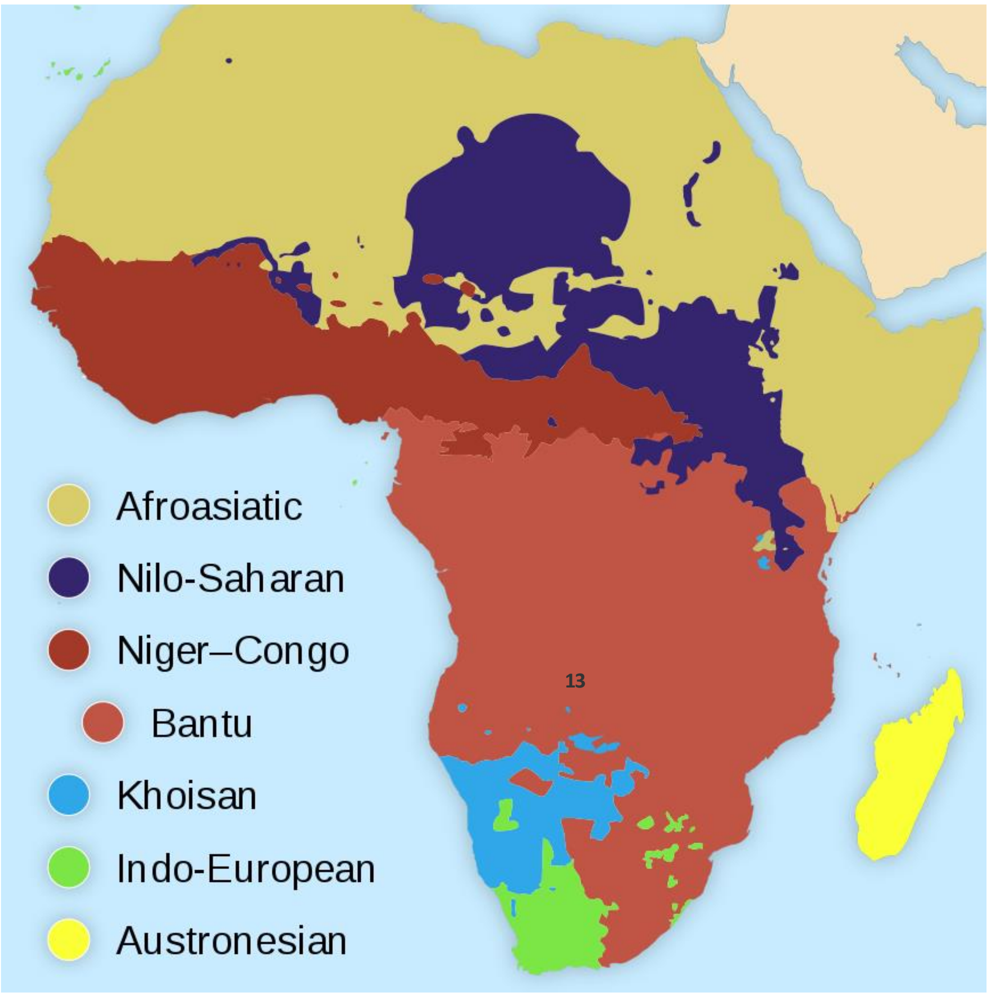
Ethnologue

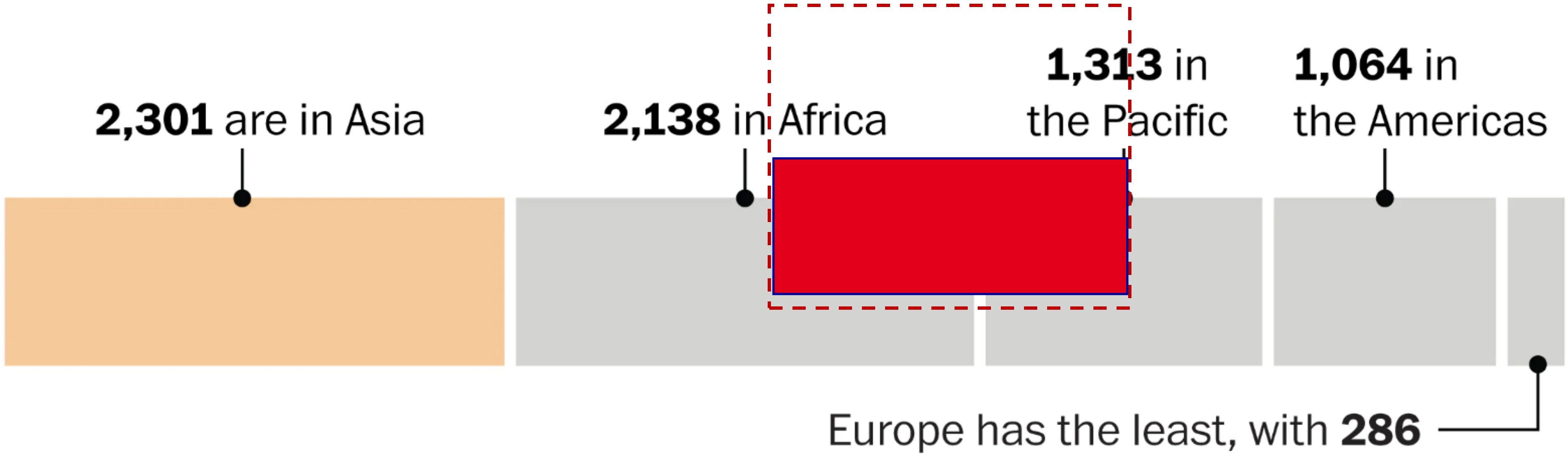


- Only 7.4% are institutional
  - often used by governments, schools, and mass media
  - even less are supported by language technologies

African languages are Low-resource Languages

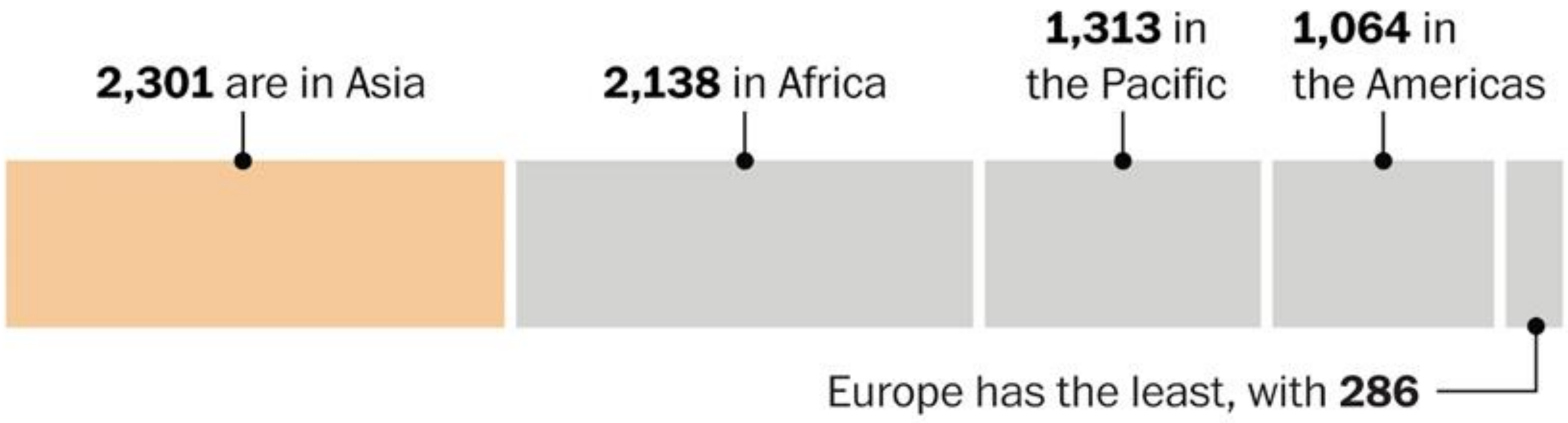
- African languages are **under-represented in NLP research** despite **spoken by 1.5B people**





Language families (Wikipedia) - under-researched

Under-Resourced Languages: NLP Research



Washington Post (2015)

- Languages are **not treated equally** by researchers
- Number of NLP publications extracted from ACL Anthology
- Fewer representation of **African**, **Asian** and **The Americas** languages

Number of NLP publications per language

Blasi et al. (2022). Systematic Inequalities in Language Technology Performance across the World's Languages. In ACL

Under-resourced languages: Labeled + Unlabeled data

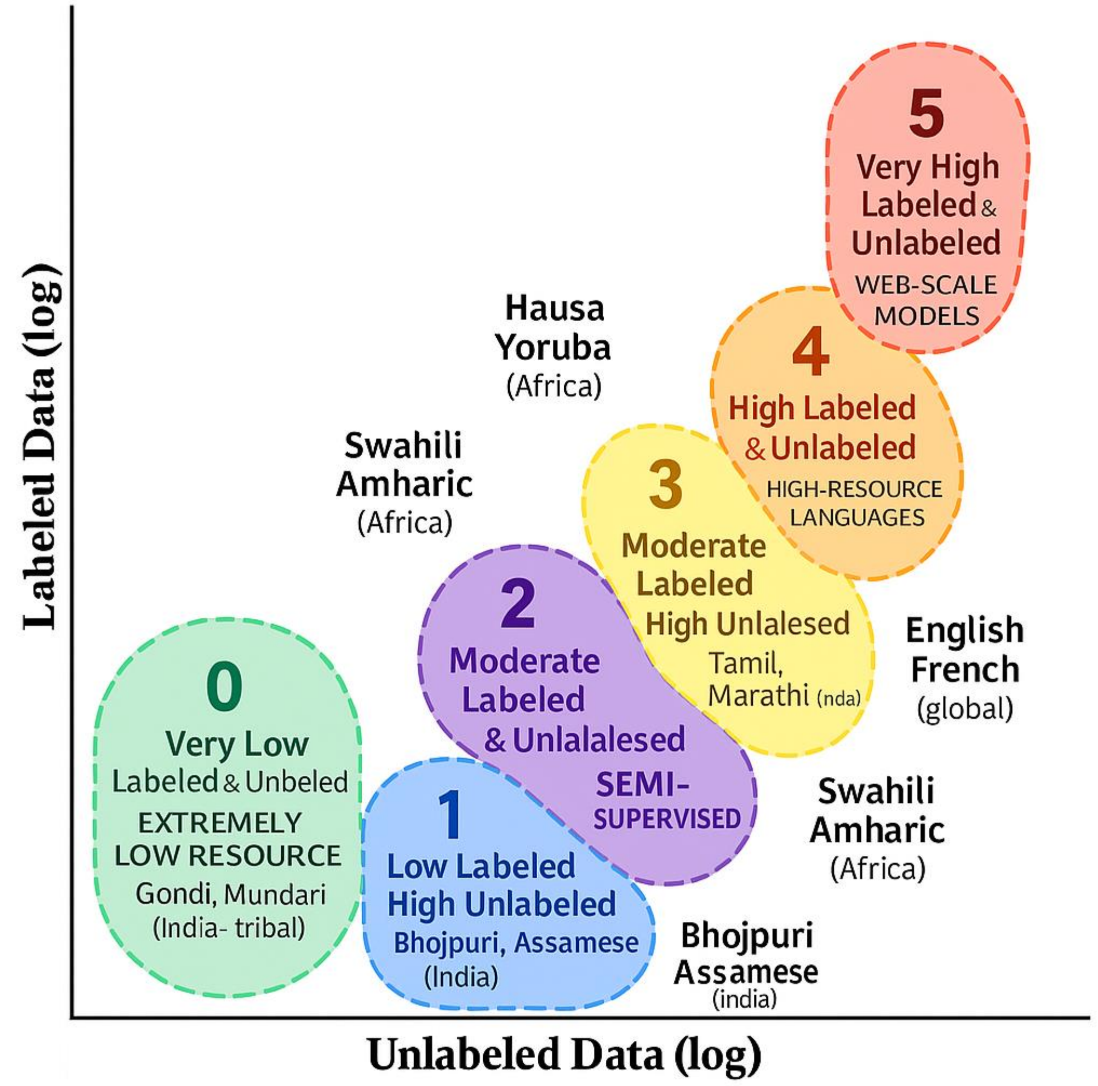
Six-class categorization of languages based on Joshi et al (2020)

- Unlabeled corpora
- Labeled corpora

Joshi 5 means high resources and
Joshi 2 is pretty low

0 – The Left-Behinds 🏚️
1 – The Scraping-Bys 🥄
2 – The Hopefuls ✨
3 – The Rising Stars 🌟
4 – The Underdogs 🐾
5 – The Winners 🏆

Source: The State and Fate of Linguistic Diversity and Inclusion in the NLP World Joshi et al (2020).

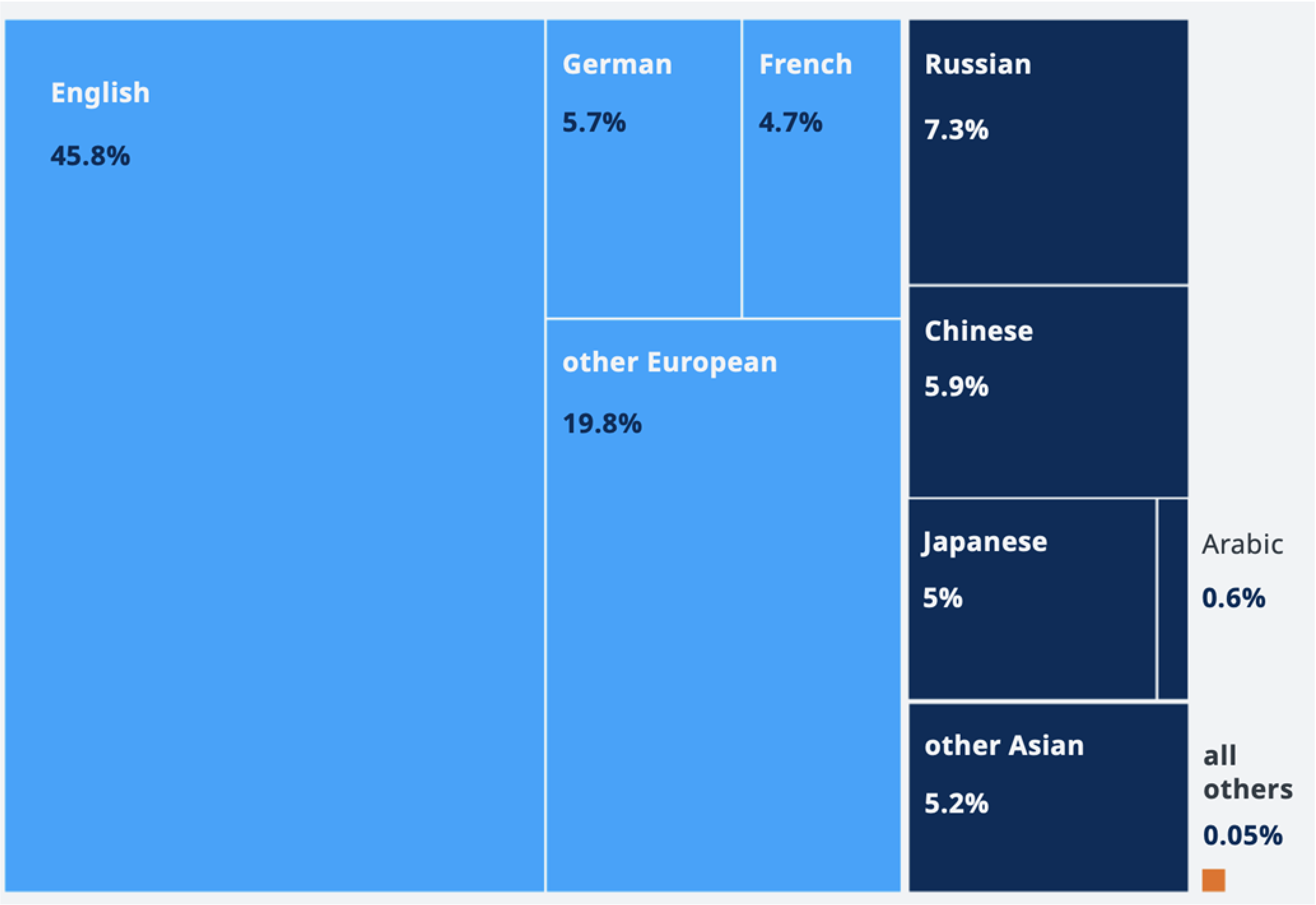


No unlabeled texts
80% of languages

The State and Fate of Linguistic Diversity and Inclusion in the NLP World

Lack of Publicly Available Dataset

Languages in the Common Crawl internet archive



Source: Common Crawl | More Info: github.com/dw-data/ai-languages

30%

World languages are African (Ethnologue)

0.05%

Low-Research: Datasets, Tools, and Benchmarks

TACL 2021

HCDS HUB OF COMPUTING & DATA SCIENCE

MasakhaNER: Named Entity Recognition for African Languages

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, Salomey Osei

Language	Sentence
English	The Emir of Kano turbaned Zhang who has spent 18 years in Nigeria
Amharic	የካኖ ኢምር በናይጀርያ ፩፰ ዓመት ያሳለፈውን ዣንግን ዋና መሪ አደረጉት
Hausa	Sarkin Kano yayi wa Zhang wanda yayi shekara 18 a Najeriya sarauta
Igbo	Onye Emir nke Kano kpubere Zhang okpu onye nke nọgoro afọ iri na asatọ na Naijịrịa
Kinyarwanda	Emir w'i Kano yimitse Zhang wari umaze imyaka 18 muri Nijeriya
Luganda	Emir w'e Kano yatikkidde Zhang amaze emyaka 18 mu Nigeria
Luo	Emir mar Kano ne orwakone turban Zhang ma osedak Nigeria kwuom higni 18
Nigerian-Pidgin	Emir of Kano turban Zhang wey don spend 18 years for Nigeria
Swahili	Emir wa Kano alimvisha kilemba Zhang ambaye alikaa miaka 18 nchini Nigeria
Wolof	Emiiru Kanó dafa kaala kii di Zhang mii def Nigeria fukki at ak juróom ñett
Yorùbá	Ẹ́míà ìlú Kánò wé láwàní lé orí Zhang ẹni tí ó ti lo ọdún méjìdínlógún ní orílè-èdè Nàìjíríà

Table 2: Example of named entities in different languages. PER, LOC, and DATE are in colours purple, orange, and green, respectively.

MasakhaNER

HCDS HUB OF COMPUTING & DATA SCIENCE

- First large, publicly available NER dataset for **10 African languages**, curated by **native speakers** from local news sources.
- High-quality annotation via collaborative workshops, achieving strong inter-annotator agreement.
- **Benchmarked** multiple NER models (CNN-BiLSTM-CRF, mBERT, XLM-R); **language-adaptive fine-tuning** boosted performance.
- Gazetteer features and transfer learning (cross-domain/ cross-lingual) improved NER for low-resource languages.
- Released data, code, and models to spur future African NLP research and applications.

https://github.com/masakhane-io/masakhane-ner/

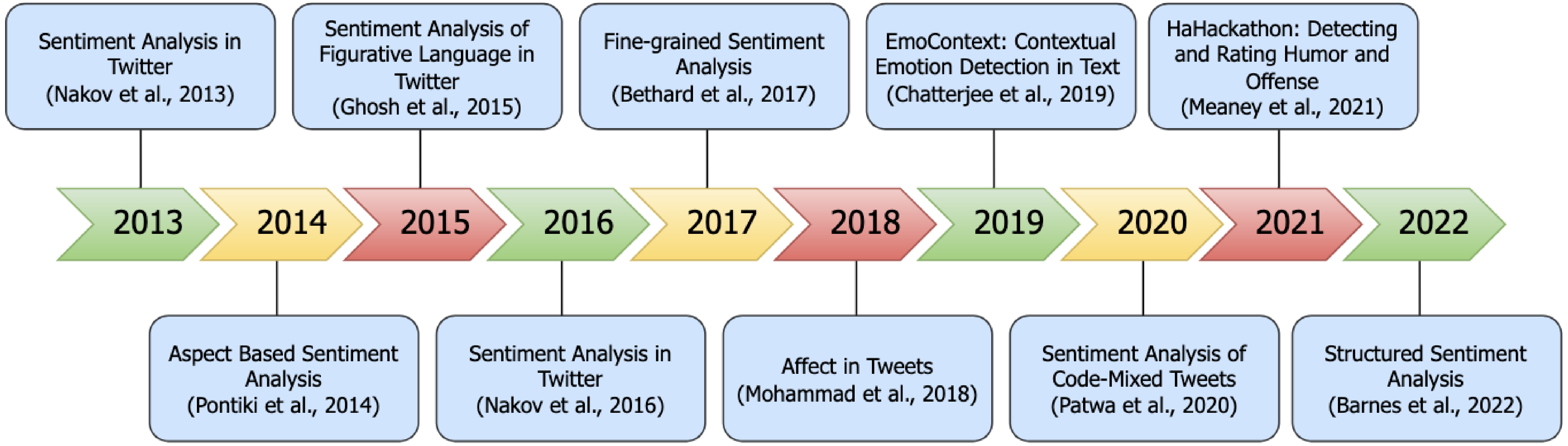
SemEval-2023 Task 12:

Sentiment Analysis for African Languages (AfriSenti-SemEval)

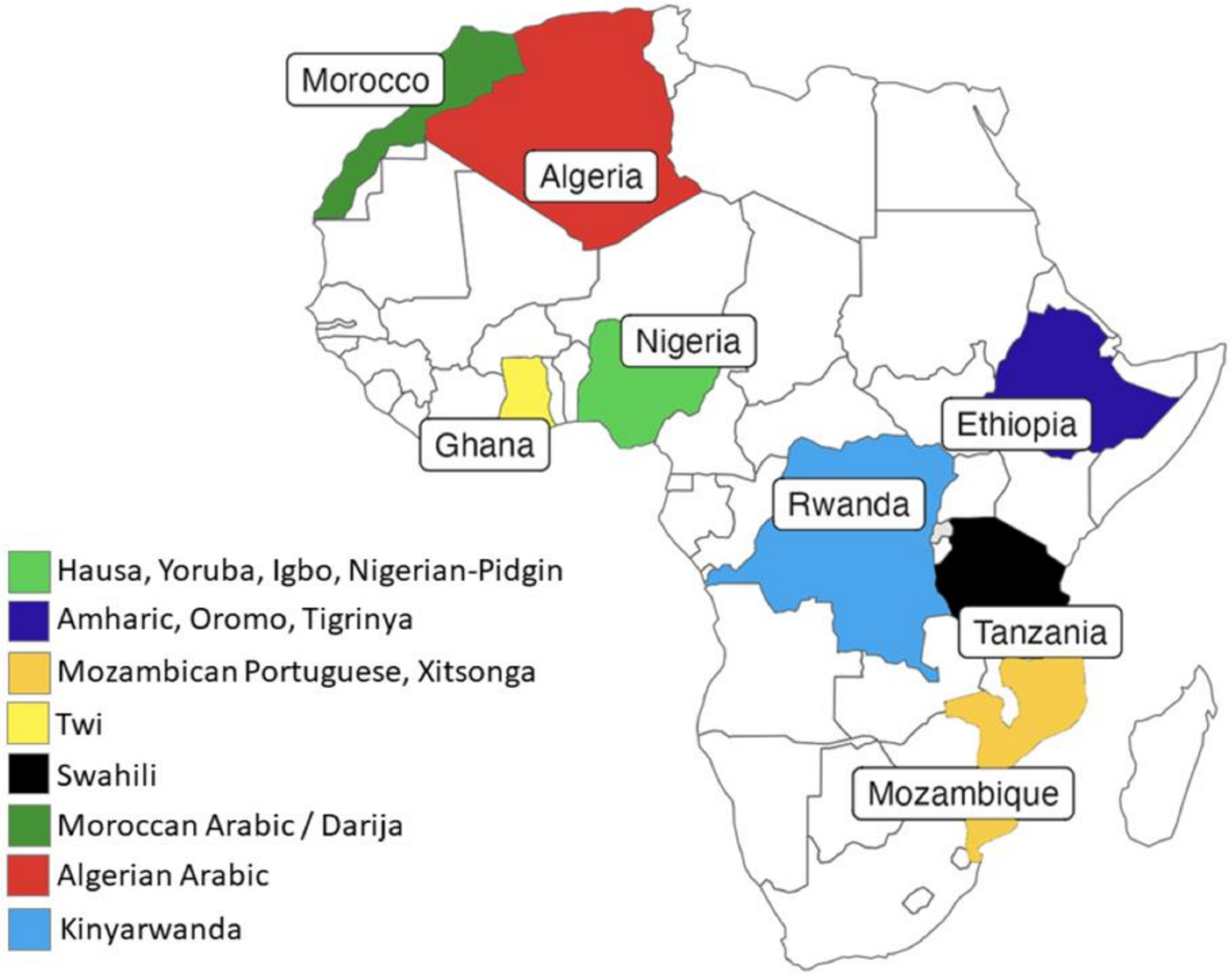
https://afrisenti-semeval.github.io/

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Nedjma Ousidhoum, Abinew Ayele, Saif M. Mohammad, Meriem Beloucif, Sebastian Ruder

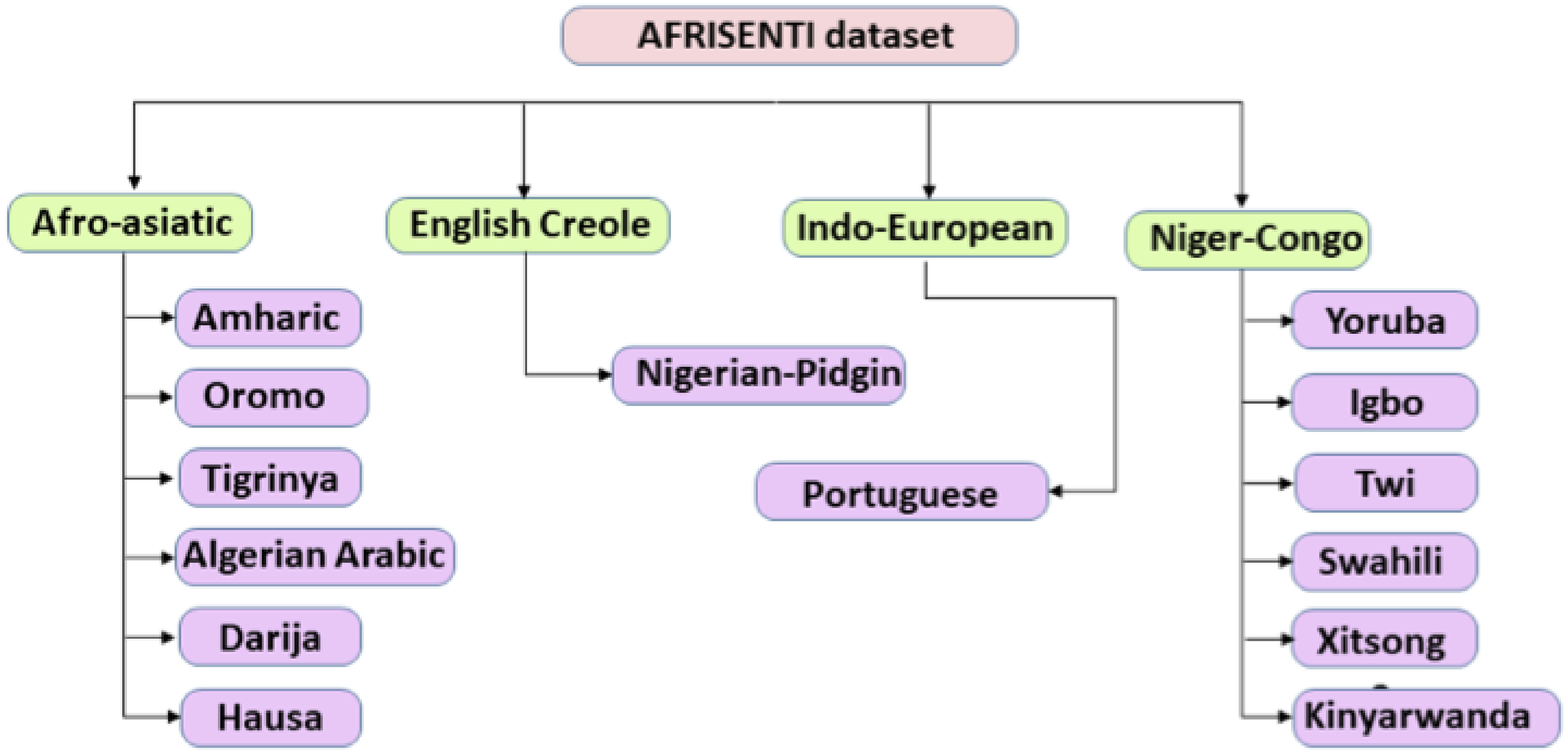
SemEval 2023, Toronto, Canada

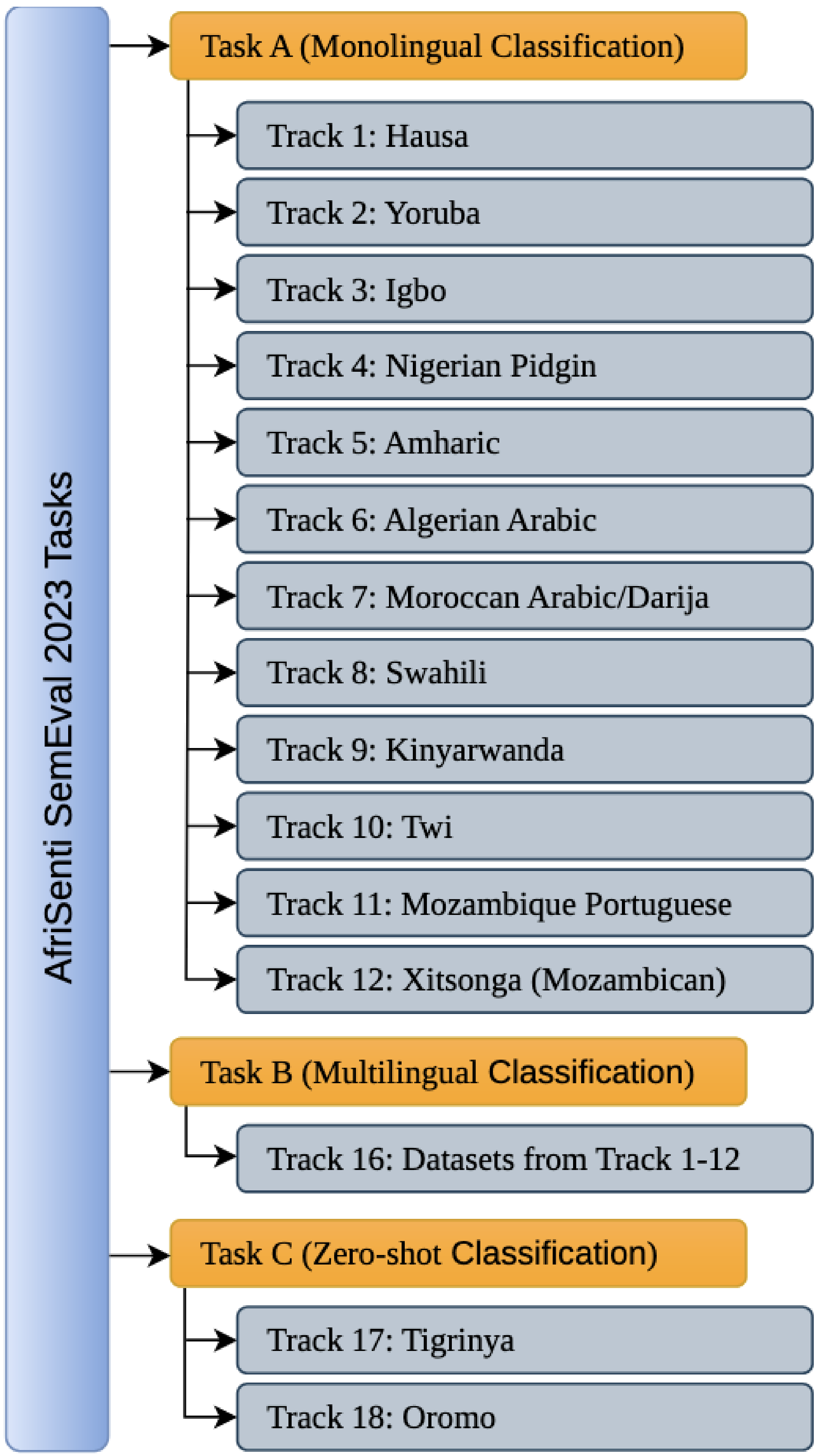


AfriSenti dataset



22





Three tasks

1. **Monolingual**: Classify the sentiment (positive, negative, neutral) of tweets in a single African language
2. **Multilingual**: Train and evaluate a sentiment classifier on a combined dataset including multiple African languages.
3. **Zero-shot**: Test a sentiment classifier trained on one language's data to predict sentiment in another (unseen) African language.

Lang.	Tweet	Sentiment
amh	ያ ጨካኝ አረመኔ ታስሮ ይኸው ካቴና ገብቶለታል ይሉናል። ቆይ አስረው የጀበና ቡና እየጋበዙት ነው እንዴ?	negative
arq	@user .... الشروق هذه من خرجت وهى تتاع تهديل، مستوى منحط وشعبوي	negative
ary	واش بغيتوهم يبداو يتكرفسو على العادي والبادي عاد تبقاو أنتما على خاطر خاطركم	negative
ary	rabi ykhali alhbiba makayn ghir nachat o chi machat	positive
hau	@USER Aunt rahma i luv u wallah irin totally dinnan	positive
ibo	akowaro ya ofuma nne kai daalu nwanne mmadu	positive
kin	@user Ariko akokanu ngo inyebebe unyujijemo sisawa wangu	negative
orm	@user Jawaar Kenya OMN haala akkamiin argachuu dandeenya	neutral
por	Honestidade é algo que não se compra. Infelizmente a humanidade esqueceu disso por causa das suas ambições.	positive
pcm	E don tay wey I don dey crush on this fine woman ...	positive
swa	Asante sana watu wa Sirari jimbo la Tarime vijijini Huu ni Upendo usio na Mashaka kwa Mbunge wenu John Heche	positive
tir	@user ከመኸረኩም እንተኸይኑ:ንሕውሓት ነዞም ውሑድ ቁጽርም አባ ምጥፋእ ይሕሸኩም!	negative
tso	@user @user Yu , tindzava ? Tsika mbangui mpfana e nita ku despro-gramara	negative
twi	messi saf den check en bp na wo kwame danso wo di twe da kor aaa na wawu	negative
yor	oníròégbè aláàdúgbò ati olójúkòkòrò	negative

Lang.	AfriBERTa large	XLM-R base	AfroXLMR base	mDeBERTa base	XLM-T base	XLM-R large	AfroXLMR large
amh	56.9	60.2	54.9	57.6	60.8	**61.8**	61.6
arq	47.7	65.9	65.5	65.7	**69.5**	63.9	68.3
ary	44.1	50.9	52.4	55.0	**58.3**	57.7	56.6
hau	78.7	73.2	77.2	75.7	73.3	75.7	**80.7**
ibo	78.6	75.6	76.3	77.5	76.1	76.5	**79.5**
kin	62.7	56.7	67.2	65.5	59.0	55.7	**70.6**
pcm	62.3	63.8	67.6	66.2	66.6	67.2	**68.7**
pt-MZ	58.3	70.1	66.6	68.6	71.3	71.6	**71.6**
swa	61.5	57.8	60.8	59.5	58.4	61.4	**63.4**
tso	51.6	47.4	45.9	47.4	**53.8**	43.7	47.3
twi	**65.2**	61.4	62.6	63.8	65.1	59.9	64.3
yor	72.9	62.7	70.0	68.4	64.2	62.4	**74.1**
AVG	61.7	61.9	63.9	64.2	64.7	63.1	**67.2**

Results

- Top-performing teams in each subtask were **not affiliated with African** institutions.
  - Despite a lack of language expertise.
  - Access to **compute-resource**; GPU, while all African teams use Google **Colab**
- This highlights the need for a **more collaborative** approach to building more effective and inclusive solutions for Africa-centric sentiment analysis.
- By sharing our insights, we aim to

AfriSenti-SemEval: Stats

Submissions

500+

System Papers

29

Co-located with ACL 2023
Toronto Canada

Can we do better?

AFRISENTI
SemEval 2023

AfriHate

"No one is born hating another person because of the colour of his skin, or his background, or his religion. People must learn to hate, and if they can learn to hate, they can be taught to love, for love comes more naturally to the human heart than its opposite." — Nelson Mandela, Long Walk to Freedom

Hate Speech and Abusive Language Datasets for African Languages

Team

Project Leading Universities

Bayero University
Kano, Nigeria

Bahir Dar University,
Ethiopia

Project Partner Organizations

Language Technology

Rewire

MasaKhane

LIAAD INESCTEC

U.PORTO

Hate in Africa

HCDS HUB OF COMPUTING & DATA SCIENCE

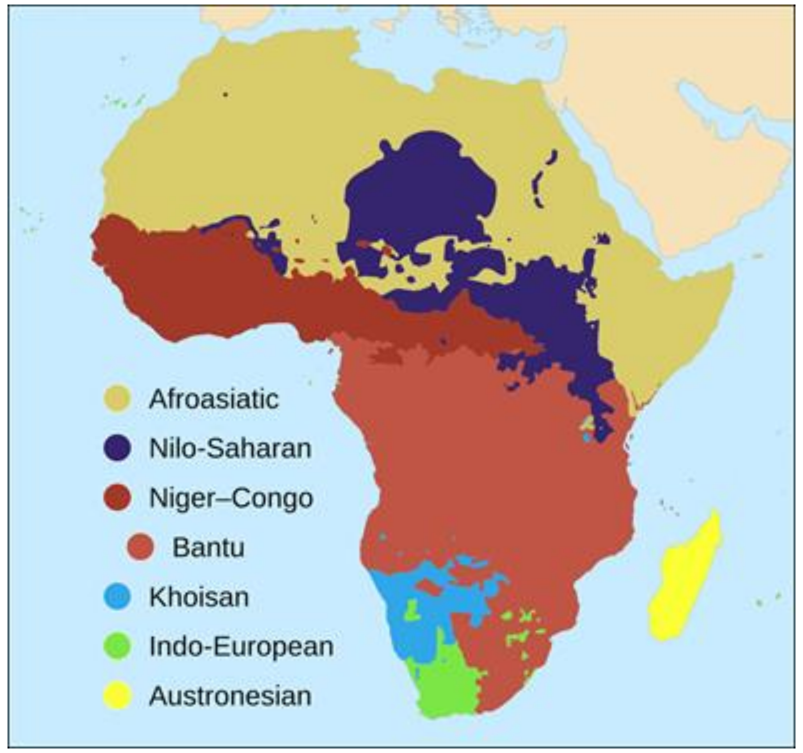
Challenges

- Online hate is a growing problem across Africa and the world.

> “Meta has failed to adequately invest in content moderation in the Global South
>
> Flavia Mwangovya

Detection is difficult

- Linguistic diversity,
- Cultural nuances, and
- Evolving online discourse.



Language families (Wikipedia)

Consequences can be dire

- Serious offline harm
- Inciting violence
- Perpetuating discrimination

ARTICLE | JULY 28, 2022

Facebook unable to detect hate speech weeks away from tight Kenyan election

NEWS

Home | InDepth | Israel-Gaza war | War in Ukraine | Climate | UK | World | Business | Politics | Culture

World | Africa | Asia | Australia | Europe | Latin America | Middle East | US & Canada

Facebook's algorithms 'supercharged' hate speech in Ethiopia's Tigray conflict

Motivation

- Language barriers
  - Lack of NLP tools for African languages.
  - Ineffective interventions (keyword-based removal) without context.
- Introducing afriHate
  - Comprehensive labeled dataset for **18 African languages**.
  - Aims to identify hate and abusive language.
- Benefits
  - Supports development of classification models for **automatic moderation**.
  - Utilizable by platform owners, **peacebuilders**, and community services.
  - Promotes NLP innovation for African languages.

AfriHate Dataset

Nigeria
Hausa, Igbo, Pidgin, Yoruba

Ghana
Twi

South Africa
isiXhosa, isiZulu

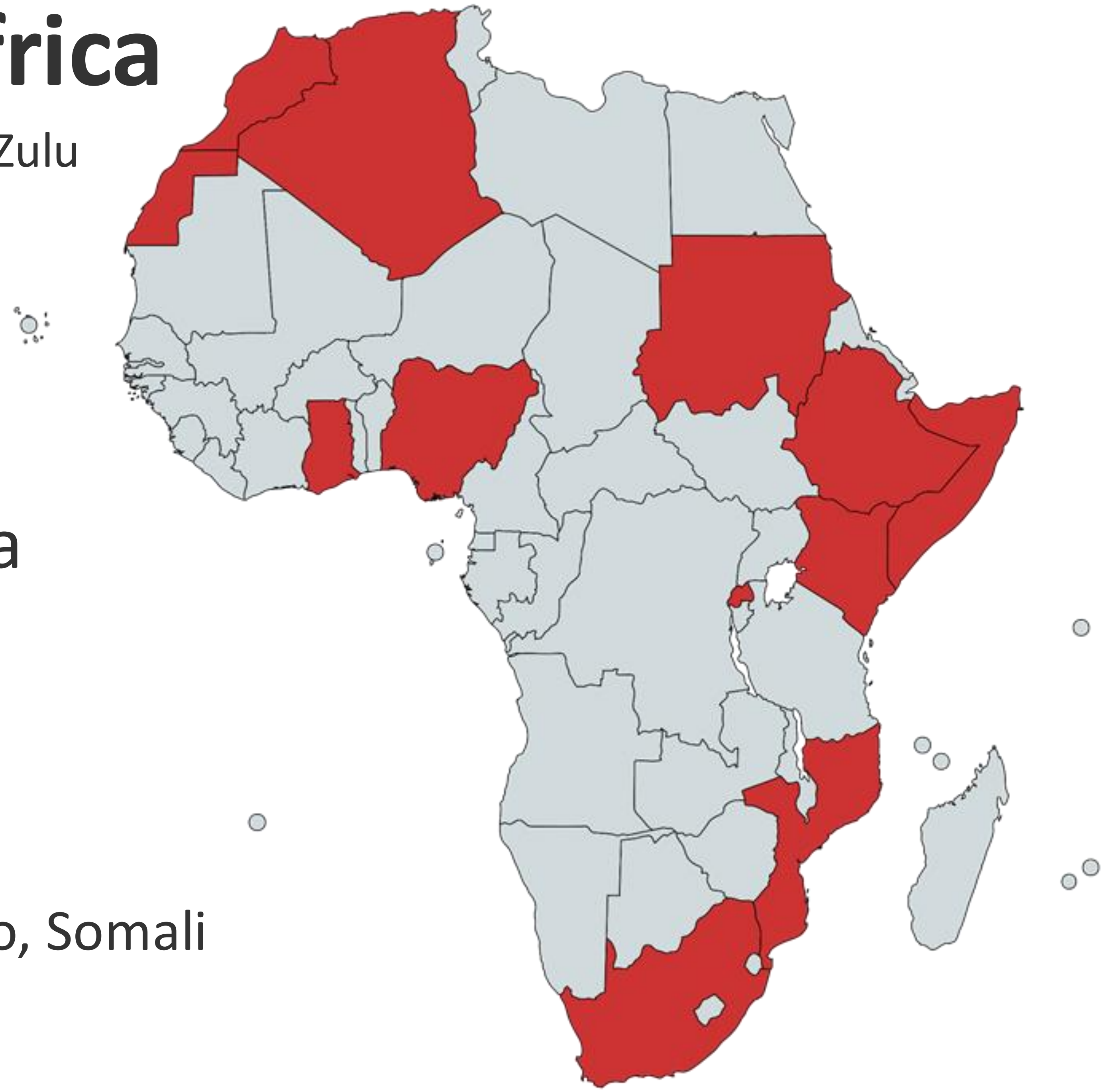
Kenya
Swahili

Morocco
Darija

Rwanda
Kinyarwanda

Algeria
Algerian Arabic

Ethiopia
Amharic, Tigrinya, Oromo, Somali



AfriHate Dataset

Annotation Advice

- Harmful content can be psychologically distressing.
  - We advised any annotator who feels anxious or uncomfortable during the process to take a break or stop the task and seek help (first point of contact was the language leads).
  - Early intervention is the best way to cope.

@USER በታሪክ ያልነበረ የኦሮሞ ግዛት እየመራ ሌላውን አይጠቅማችሁም አይናችሁን ጨፍኑ አይሰራም ... ፤ ክልል ነው። ኦሮሚያ ከ5 ይከፈል። ለአፍሪቃ ቀንድ ሥጋት ነው።

Translation

@USER While leading an oromo state that doesn't exist in history, it is not possible to cheat other who ask for regional state structure. Gurage is a regional state. Oromia should be divided into 5. It is a threat to the Horn of Africa.

What is the text category?

1

- ( ) Offensive
- (•) Hate
- ( ) Normal
- ( ) Indeterminate

How hate is this tweet?

2

Very Hate ( ) ( ) ( ) ( ) (•) Less Hate

What is the target of the hate?

3

- [x] Ethnicity
- [ ] Religion
- [ ] Disability
- [ ] Gender
- [x] Politics

Others

E.g. racism, sexual orientation, etc.

Previous Submit

HCDS HUB OF COMPUTING & DATA SCIENCE

Universität Hamburg DER FORSCHUNG | DER LEHRE | DER BILDUNG

AfriHate Models

Model Development Strategies

- **Fine-tuning PLMs**
  - AfroXLMR, AfriBERTa, AfriTeVa
- **Zero and Few-shot Learning**
  - SetFit
- **Zero- and Few-shot Prompting of LLMs**
  - GPT-4o (closed model)
  - InbubaLM (SLM)
  - mT0-small
  - BLOOMZ 7B
  - Mistral
  - Aya-23-35B
  - LLaMa 3.1
  - Gemma

HCDS HUB OF COMPUTING & DATA SCIENCE

AfriHate Models

Results

model	amh	ary	arq	hau	ibo	kin	oro	pcm	som	tir	twi	xho	yor	zul	avg.
Monolingual															
AfriBERTa	69.54	67.93	30.48	82.28	89.53	79.43	73.43	66.90	65.52	73.07	74.54	81.07	72.37	83.75	72.33
AfriTeVa V2	73.91	76.71	25.25	79.06	83.95	77.60	71.61	68.69	69.65	72.36	64.96	54.67	**79.88**	69.05	68.73
AfroXLMR	70.65	80.16	61.18	81.93	89.30	**80.72**	72.11	67.98	66.84	74.52	77.17	82.49	72.15	83.44	76.15
AfroXLMR-76L	74.36	80.05	53.52	**82.78**	89.59	79.58	76.63	68.38	71.09	76.27	76.65	84.40	72.35	84.65	76.45
Multilingual															
AfroXLMR-76L	**75.25**	**80.76**	**63.31**	82.20	**89.85**	79.56	**77.62**	**69.20**	**72.26**	**77.55**	**78.68**	**86.83**	74.32	**86.81**	**78.16**

Universität Hamburg DER FORSCHUNG | DER LEHRE | DER BILDUNG

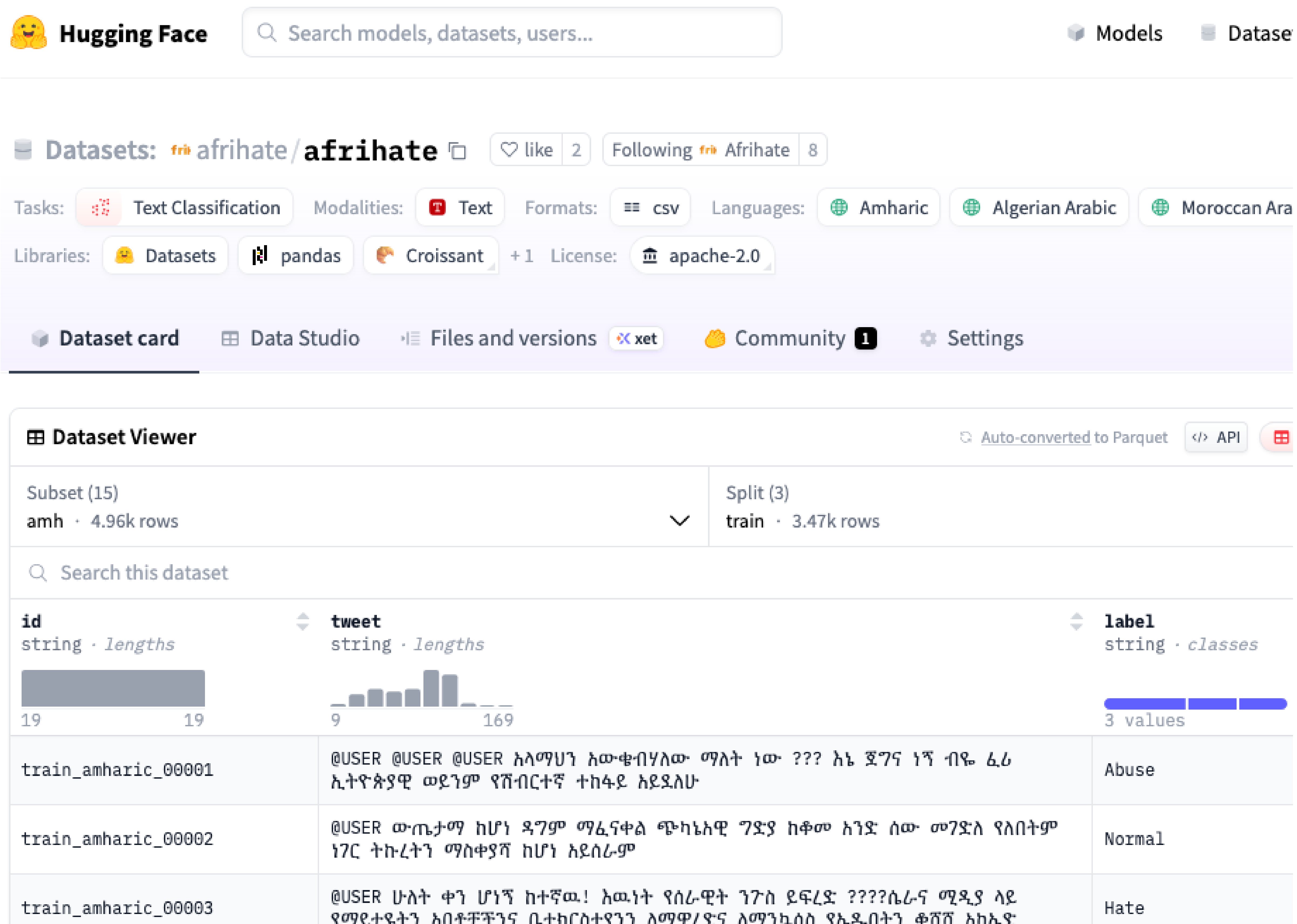
AfriHate

Ethical Concerns

- Censorship and freedom of speech
- Curbing hate speech vs. Preserving free expression
- Ensuring transparent and accountable moderation policies

Dataset – Hugging Face

https://huggingface.co/datasets/afrihate/afrihate



SemRel 2024:

A Collection of Semantic Textual Relatedness Datasets for 13 Languages

https://semantic-textual-relatedness.github.io

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine De Kock, Genet Shanko Dekebo, Oumaima Hourrane, Gopichand Kanumolu, Lokesh Madasu, Samuel Rutunda, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Hailegnaw Getaneh Tilaye, Krishnapriya Vishnubhotla, Genta Winata, **Seid Muhie Yimam**, Saif M. Mohammad

Semantic Textual Relatedness (STR)

- STR involves:
  - Semantic Textual Similarity (**STS**).
- All **commonalities** between two units of text (sentences):
  - Sentences **on the same topic.**
  - Sentences expressing **the same view.**
  - Sentences originating from **the same time period.**
  - Sentences **elaborating on** (or **following**) the other.
  - …

Semantic Textual Relatedness (STR)

Pair 1	There was a lemon tree next to the house	I have a green hat
Pair 2	I am feeling sick	Get well soon

- Most people will agree that the sentences **in pair 2** are **more related** than the sentences in **pair 1**.
- Most people will also agree that the sentences in pair 2 are related but **not similar**.

STR Data Creation

Sentence Pairing

Random selection results in many unrelated sentences.

We use heuristics to ensure a sufficient number of instances for each band of relatedness.

(High, medium, low, or unrelated).

STR Data

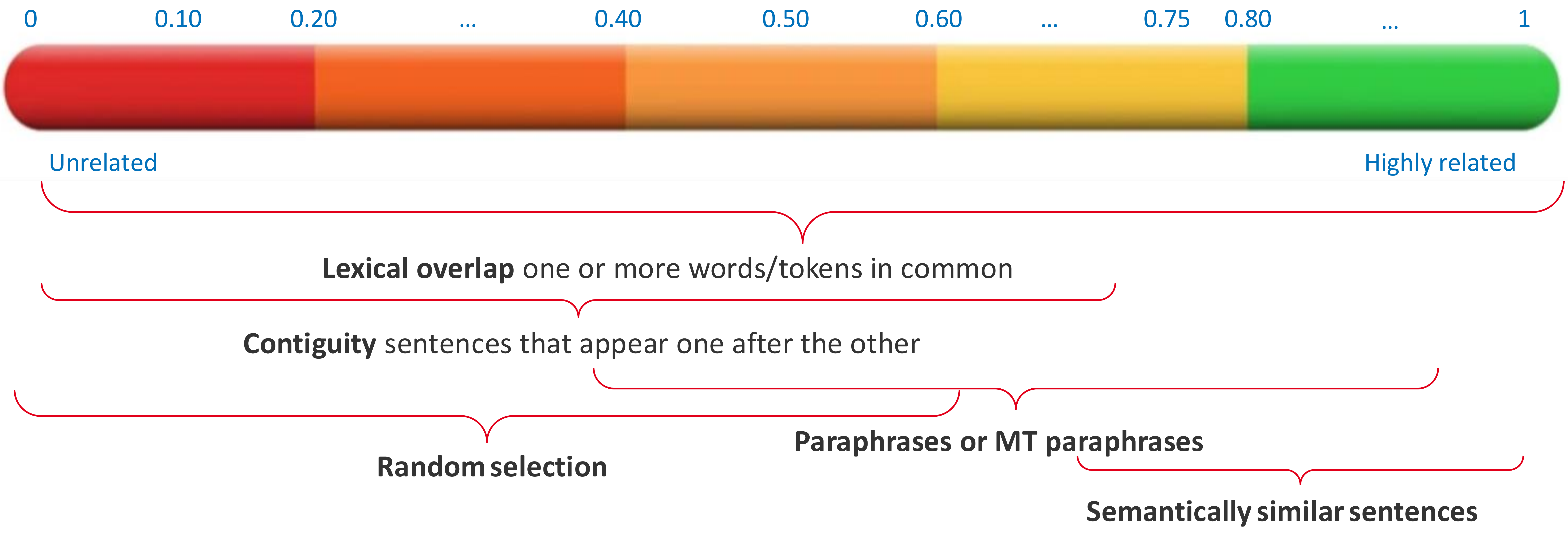
- **Related** and **unrelated** do not have clear boundaries.

- We use comparative annotations: **Best-Worst Scaling (BWS).**

STR Data Creation

Sentence Pairing Heuristics

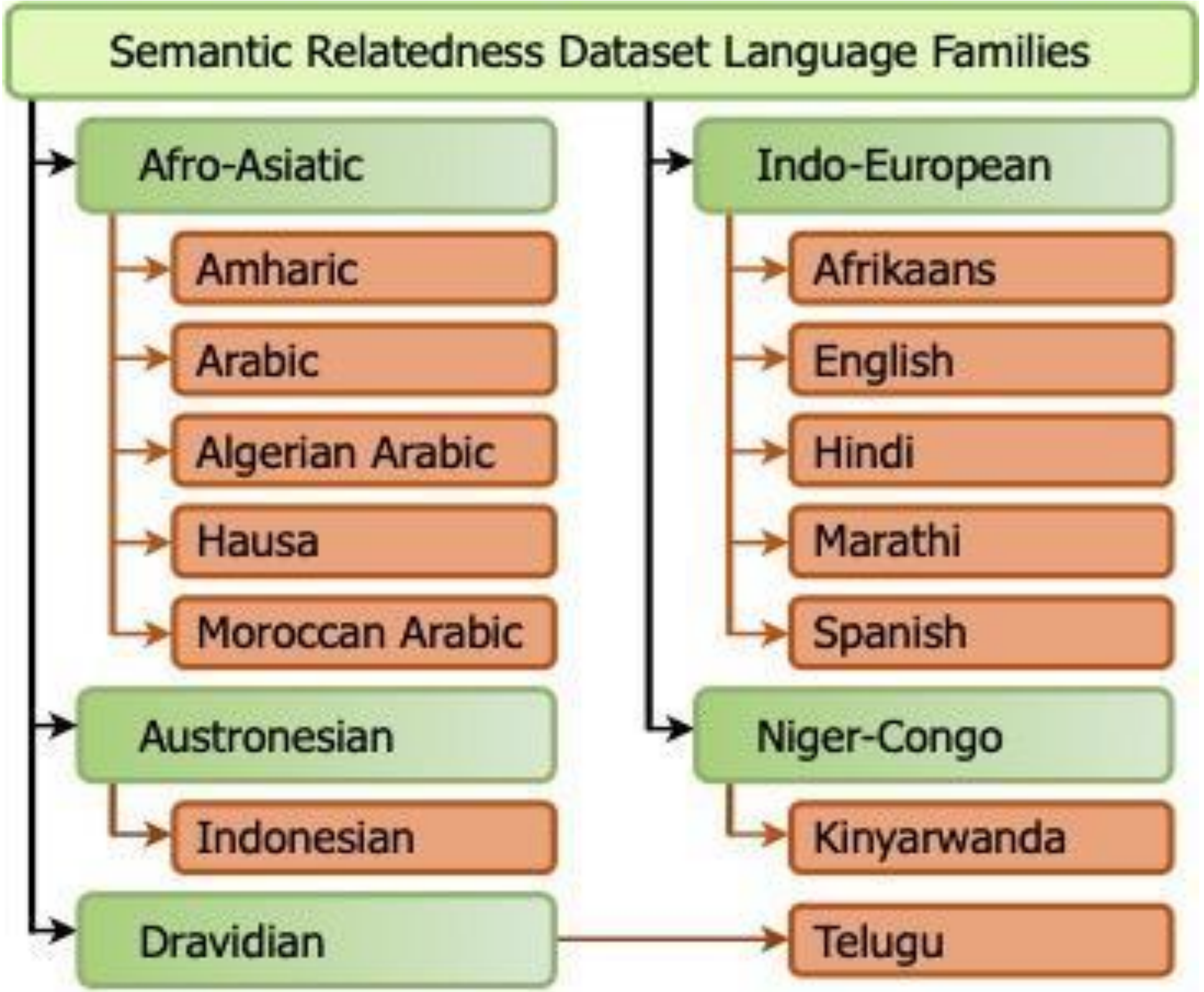
We build datasets within a wide range of relatedness scores.



STR Data Creation

Languages

13 languages from 5 language families



STR Data Creation

Data Annotation using BWS

≥ 3

That's difficult. They're both great That's really hard they are both great!	👍
That's difficult. I think it’s easy.	
There is a lemon tree next to the house. I love reading next to the lemon tree.	
I was travelling. She bought a new phone.	👎

We generate real-valued scores based on **the number of times a pair was chosen as best** and **the number of times it was chosen worst.**

STR Data Creation

Data Instances

L	Sentence #1	Sentence #2	Score
Eng	If that happens, just pull the plug.	If that ever happens, just pull the plug.	1.0
Hau	Haka ya furta a cikin jawabin sa na murnar cikar Najeriya shekaru 61 da samun ‘yanci.	Ya yi wannan iƙirarin e a cikin jawabin sa na murnar cikar Najeriya 61 da samun ‘yanci a ranar Juma’a.	0.94
Amh	መግለጫውን የተከታተለው የአዲስ አበባው ዘጋቢያችን ሰሎሞን ሙጬ ዝርዝር ዘገባ አለው ።	በስፍራው ተገኝቶ የተከታተለው የአዲስ አበባው ዘጋቢያችን ሰሎሞን ሙጬ ያጠናቀረውን ልኮልናል ።	0.88
Ind	Pendidikan Desa Pusaka memiliki 4 sekolah.	Pendidikan Desa Serumpun Buluh memiliki 4 sekolah.	0.83
Arb	في الواقع، هذه المادة التي ترون واضحة وشفافة.	مركبات هذه المادة هي فقط الماء والبروتين	0.78
Ary	درجة فهاد المناطق 37الحرارة غادي تبدا بـ ..وجدو راسكوم لرمضان	40الحرارة غادي تبدا وغادي توصل لـ ..غير خرج رمضان وهي تشعل درجة فهاد المناطق	0.75
Tel	క్రికెట్ అన్ని ఫార్మాట్స్‌కు మలింగ గుడ్‌బై	కొలంబో: శ్రీలంక సీనియర్ పేసర్ లసిత్ మలింగ క్రికెట్ అన్ని రకాల ఫార్మాట్స్‌కు గుడ్‌బై చెప్పాడు.	0.62

Experiments

- Given sentence pairs, automatically determine relatedness scores.

- We assess how well system-predicted rankings of test instances aligned with human judgments.

- Metric Spearman rank correlation coefficient.

Experiments

Settings

- **Supervised settings**
  - Train using the labeled training data.
- **Unsupervised settings**
  - Train without using any labeled STS or STR datasets between text >2 words long in any language.
- **Crosslingual settings**
  - Train without using any labeled STS or STR datasets in the target language.
  - Train using labeled datasets from 1 other language.
    - I.e., English for all non-English datasets and Spanish for the English dataset.

- **Note: Datasets without training sets (afr, arb, hin, ind) were only used in unsupervised and crosslingual settings.**

Experiments

Models

- **Baseline**
  - **Lexical Overlap** number of unique unigrams occurring in sentences.
- **Supervised**
  - **Multilingual** mBERT and XLMR for unsupervised settings.
  - **Monolingual** Language-specific LMs (e.g., BERTO, IndicBERT, DziriBERT, etc.).
- **Unsupervised and Crosslingual**
  - LaBSE.

Results

		afr	amh	arb	arq	ary	eng	esp	hau	hin	ind	kin	mar	tel
Baseline	**Overlap**	0.71	0.63	0.32	0.40	0.63	0.67	0.67	0.31	0.53	0.55	0.33	0.62	0.70
Unsupervised	**mBERT**	0.74	0.13	0.42	0.37	0.27	0.68	0.66	0.16	0.62	0.50	0.12	0.65	0.66
	XLMR	0.56	0.57	0.32	0.25	0.17	0.60	0.69	0.04	0.51	0.47	0.13	0.60	0.58
Supervised	**LaBSE**	-	0.85	-	0.60	0.77	0.83	0.70	0.69	-	-	0.72	0.88	0.82
Crosslingual	**LaBSE**	0.79	0.84	0.61	0.46	0.80	0.62	0.62	0.76	0.47	0.67	0.57	0.84	0.82

HCDS HUB OF COMPUTING & DATA SCIENCE

SemEval 2025-Task 11: Bridging the Gap in Text-Based Emotion Detection

[1]Shamsuddeen Hassan Muhammad*, Nedjma Ousidhoum*, Idris Abdulmumin, **Seid Muhie Yimam**,
Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay,
Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali,
Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou,
Saif M. Mohammad

https://github.com/emotion-analysis-project/SemEval2025-Task11

SemEval 2025 Task 11: Text-Based Emotion Detection

Human communication

is deeply emotional

Multilingual and cultural challenges

Emotional expression varies across languages, cultures, and contexts (perceived subjectively).

Downstream applications

Mental health, social media monitoring, dialogue systems, etc.

SemEval 2025 Task 11: Text-Based Emotion Detection

Focuses on **perceived emotions**

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

32 Languages

Dataset

Emotion labels

Anger, Disgust, Sadness, Joy, Fear, Surprise, Neutral

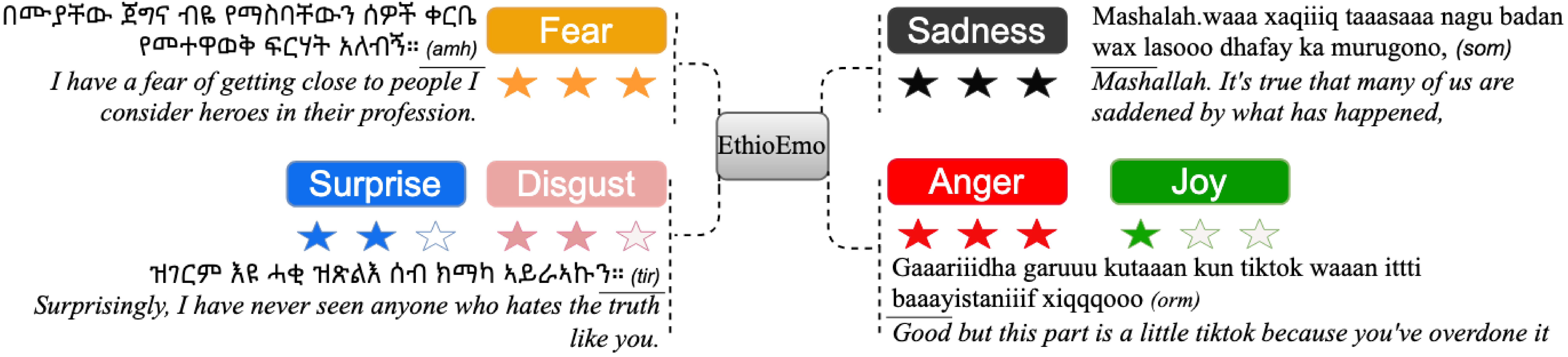
Emotion intensity

0 → no emotion

1 → low intensity

2 → moderate intensity

3 → high intensity



Task Setup

- **Track A (Multi-label Emotion Detection)**
  - **Classes:** *joy, sadness, fear, anger, surprise, and disgust*
- **Track B (Emotion Intensity Detection)**
  - **Classes:** 0, 1, 2, or 3
- **Track C (Cross-lingual Emotion Detection)**
  - **Classes:** *joy, sadness, fear, anger, surprise, and disgust*

Evaluation Metrics

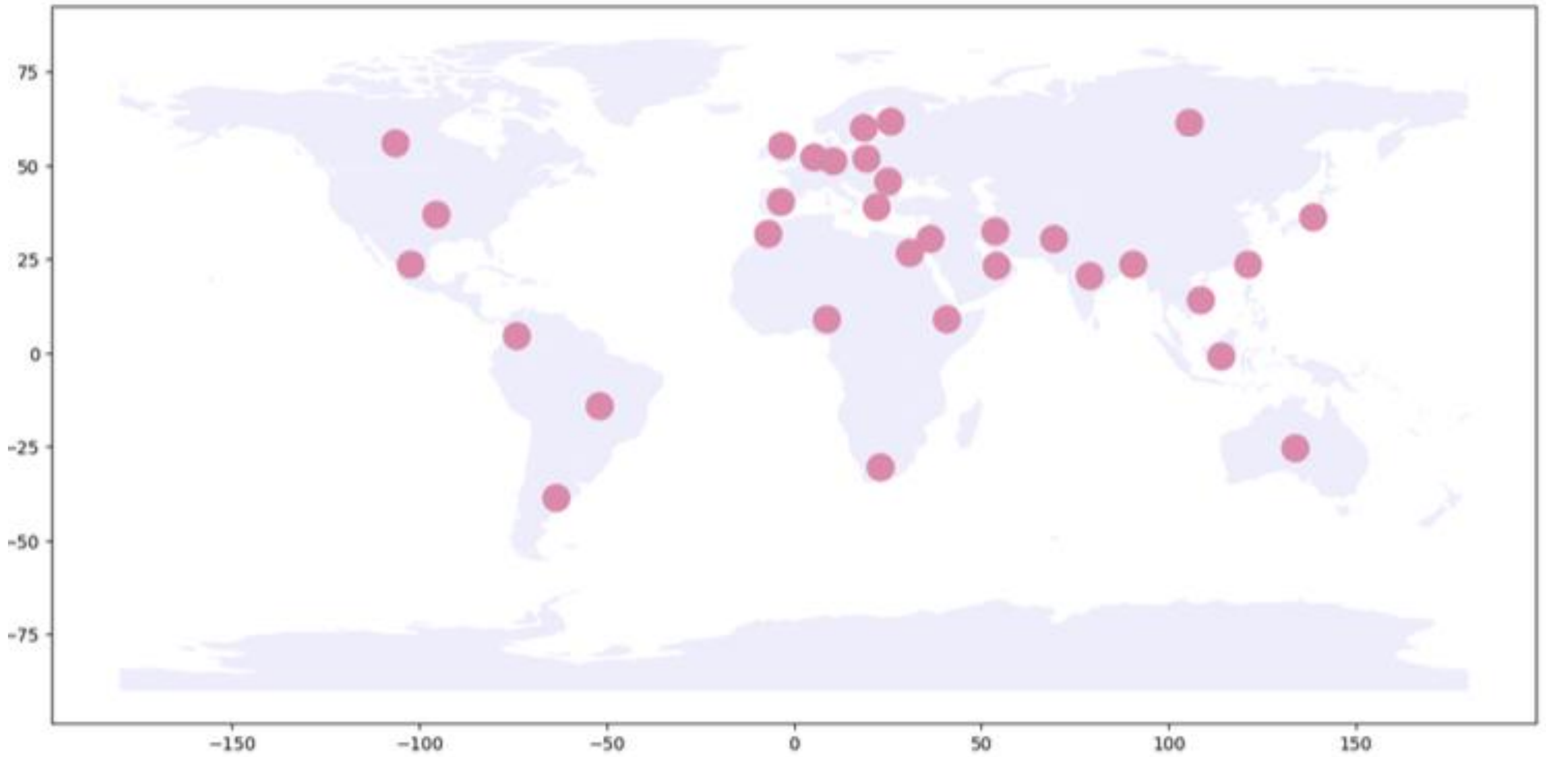
1. Average macro F-score
2. Pearson correlation coefficient

Baseline

1. Majority class
2. Fine-tuned RoBERTa

Tasks Summary

HCDS HUB OF COMPUTING & DATA SCIENCE



700+
Registered Participants

362
Submitted system during evaluation phase

93
Submitted system description paper

220

Task A: Multi-label Emotion Detection

96

Task B: Emotion Intensity Detection

46

Task C: Cross-lingual Emotion Detection

Takeaways: Popular Methods

HCDS HUB OF COMPUTING & DATA SCIENCE

- Most top-performing teams favored fine-tuning and prompting **LLMs**
- **Full fine-tuning** and **parameter-efficient fine-tuning** were the most commonly used strategies to enhance performance
- For prompting, **few-shot**, **zero-shot**, and **chain-of-thought** prompting were the most frequently used techniques.
- Traditional transformer-based models, particularly **XLM-RoBERTa**, **mBERT**, **DeBERTa**


ACL

2025 Annual Conference of the
Association for Computational Linguistics

The 19th Workshop on Semantic Evaluation (Semeval-2025)

BEST TASK AWARD

Presented to:

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Lima Ruas, Meriem Beloucif, Christine de Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Dario Mario Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou and Saif M. Mohammad

SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection

August 01, 2025

Aiala Rosa, Sara Rosenthal, Marcos Zampieri, Debanjan Ghosh
On Behalf of SemEval Workshop Organizers

SIGLEX
ACL Special Interest Group

HCDS
HUB OF COMPUTING
& DATA SCIENCE

A Case Against Implicit Standards: Homophone Normalization in Machine Translation for Languages that Use the Ge'ez Script

Hellina Hailu Nigatu[1], Atnafu Lambebo Tonja[2], Henok Biadglign Ademtew[3], Hizkel Mitiku Alemayehu[4], Negasi Haile Abadi[5], Tadesse Destaw Belay[6], Seid Muhie Yimam[7]

[1]UC Berkeley, [2] MBZUAI, [3] Vella AI, [4] Paderborn University, [5] Lesan AI, [6]Instituto Politécnico Nacional, [7]University of Hamburg

Correspondence: hellina_nigatu@berkeley.edu

EMNLP 2025
Suzhou, China | 中国苏州

Rethinking Homophone Handling

- **Traditional Normalization:**
  - Applies pre-processing during training, standardizing spelling, often at the cost of linguistic variance.

Homophones: <ዓ> and <አ> ➔ ⟨a⟩.

Eye = ‘ዓይን’ Or ‘አይን’

- **Our Innovation:**
  - Post-inference normalization: Mapping characters after model predictions, not during training.
  - Advantages:
    - Preserves dialectal, stylistic, and spelling variability.
    - Maintains model generalization to language variation.
    - Improves metric scores without sacrificing natural language features.

- **Goal: Foster more inclusive, language-aware NLP systems that respect linguistic diversity.**

Major Findings on Homophone Normalization

- Normalization improves automatic scores but **reduces linguistic** and **dialectal** diversity.
- It harms **transfer learning** across related languages (e.g., Tigrinya, Ge’ez).
- Post-inference normalization slightly **boosts scores** while preserving variation.
- Embedded standards in training models influence **language flexibility** and model behavior.

Modelling Language and Low-Resource Data

Semantic models and benchmark datasets for Amharic

Seid Muhie Yimam and Abinew Ali Ayele and Gopalakrishnan Venkatesh
and Ibrahim Gashaw and Chris Biemann

HCDS HUB OF COMPUTING & DATA SCIENCE

Pre-processing

- Sentence segmenter
- Word tokeneizer
- Text normalizer
- Text romanizer

Installation

```
pip install amseg
```

Segmentation & Tokenization

```
from amseg.amharicSegmenter import AmharicSegmenter
sent_punct = []
word_punct = []
segmenter = AmharicSegmenter(sent_punct,word_punct)
words = segmenter.amharic_tokenizer("አበበ በሶ በላ።")
sentences = segmenter.t("አበበ በሶ በላ። ከበደ ጆንያ ተሸከመ!ለምን?"))
```

words = ['አበበ', 'በሶ', 'በላ', '።']

sentences = ['አበበ በሶ በላ።', 'ከበደ ጆንያ ተሸከመ!', 'ለምን?']

Normalization & Romanization

```
normalized = normalizer.normalize('ሑለት ሦስት')
romanized = romanizer.romanize('ሑለት ሦስት')
```

normalized = 'ሁለት ሶስት'

romanized = 'ḥulat śosət'

Corpus & Dataset

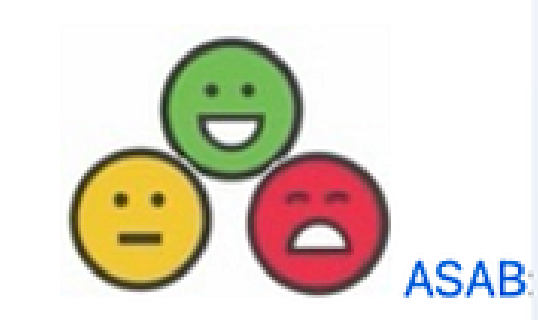
- Around **6.5m** sentences of free text
- POS tagging dataset of **35k** sentences
- Named entity recognition dataset of size **4.2k** sentences
- Around **9.4k** tweets

Mendeley Data

This is your published dataset

Amharic corpus

https://data.mendeley.com/datasets/dtywyf3sth/1

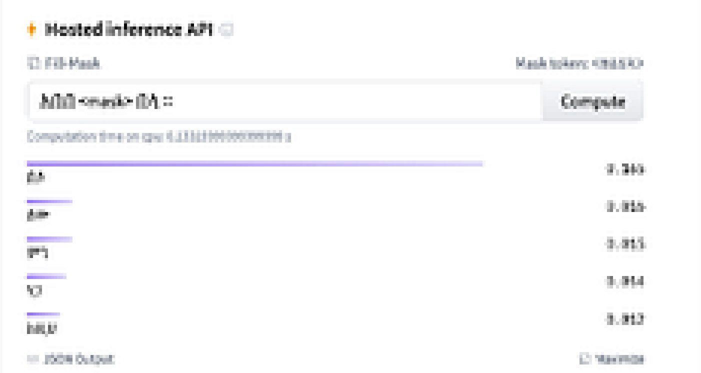


https://github.com/uhh-lt/ASAB

Semantic Models

- RoBERTa
- FLAIR
- FastText
- Word2Vec

AmRoBERTa



https://huggingface.co/uhhlt/am-roberta

word2Vec

loading the models

```
In [4]: import gensim
        model = gensim.models.KeyedVectors.load_word2vec_format('model
In [5]: model.most_similar('በሬ')
Out[5]: [('ሙክት', 0.6236802339553833),
         ('በግ', 0.5883780000183105),
         ('ወይፈን', 0.5801944732066016),
         ('በበሬ', 0.5740243196487427),
         ('በሬውን', 0.5653303861618042),
         ('በሬው', 0.5652252435684204),
         ('አህያ', 0.5641633272171021),
         ('ዓበሬ', 0.5546498298645021),
         ('ኮርማ', 0.5448639988899231),
         ('ሰንጋ', 0.5395880901336671)]
```

Tasks

- POS tagger
- NER
- Sentiment

Amharic NER model

```
from flair.data import Sentence
from flair.models import SequenceTagger

# load the model you trained
model = SequenceTagger.load(am_ner_model)
# create example sentence
sentence = Sentence('አበበ በሶ በላ ።')

# predict tags and print
model.predict(sentence)

print(sentence.to_tagged_string())
```

አበበ <B-PER> በሶ በላ ።

Amharic POS tagger

```
from flair.models import SequenceTagger
classifier = SequenceTagger.load(am_pos_model)

# create example sentence
sentence = Sentence('አበበ ብዙ በሶ በላ ።')
# predict class and print
classifier.predict(sentence)

print(sentence.to_tagged_string())
```

አበበ <N> ብዙ <ADJ> በሶ <N> በላ <V> ። <PUNC>

https://github.com/uhh-lt/amharicmodels

AfroXLMR-Social: Adapting Pre-trained Language Models for African Languages Social Media Text

Tadesse Destaw Belay[1], Israel Abebe Azime[2], Ibrahim Said Ahmad[3,4], David Ifeoluwa Adelani[5,6], Idris Abdulmumin[7], Abinew Ali Ayele[8], Shamsuddeen Hassan Muhammad[4,9], Seid Muhie Yimam[10]

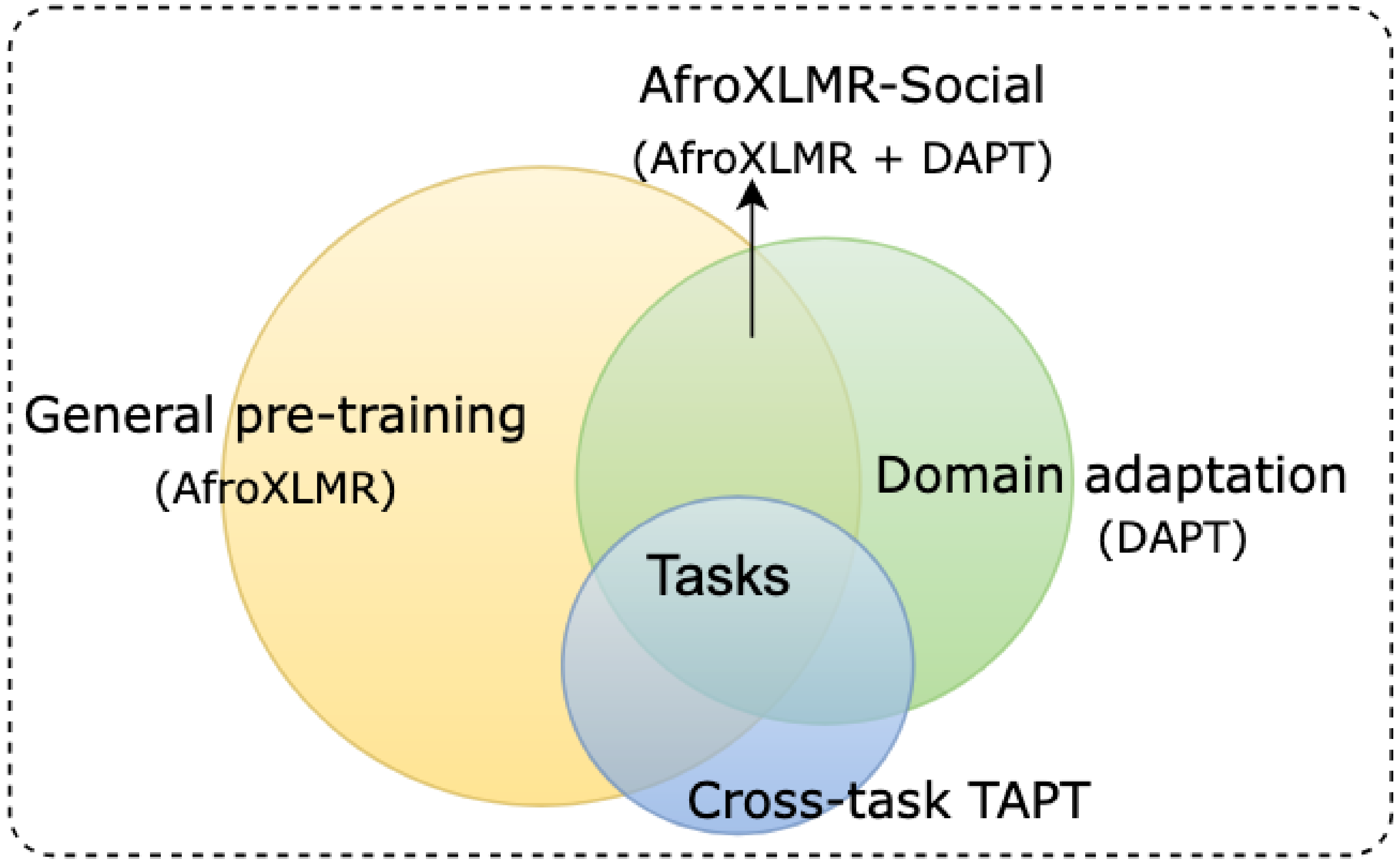
[1]Instituto Politécnico Nacional, [2]Saarland University, [3]Northeastern University, [4]Bayero University Kano, [5]Mila-Quebec AI Institute, McGill University, [6]Canada CIFAR AI Chair, [7]University of Pretoria, [8]Bahir Dar University, [9]Imperial College London, [10]University of Hamburg,

Contact: tadesseit@gmail.com

EMNLP 2025
Suzhou, China | 中国苏州

Contributions

- Introduce **AfriSocial**: a new social-domain corpus for 14 African languages (X and news data)
- Analyze domain/task adaptive continual pretraining for subjective NLP tasks in low-resource languages
- Achieve state-of-the-art results and release corpus and **AfroXLMR-Social** models to the public



	AfriSenti			AfriEmo			AfriHate	
Language	**AfroXLMR**	**+DAPT**	**Language**	**AfroXLMR**	**+DAPT**	**Language**	**AfroXLMR**	**+DAPT**
amh	50.09	**57.22**	afr	43.66	**44.57**	amh	73.54	**78.57**
arq	52.22	**64.62**	amh	68.97	**71.67**	arq	43.41	**45.96**
ary	52.86	**62.34**	ary	47.62	**52.63**	ary	75.13	**75.6**
hau	79.34	**81.66**	hau	64.30	**70.74**	hau	**81.55**	80.78
ibo	76.92	**79.8**	ibo	26.27	**54.54**	ibo	82.78	**88.05**
kin	70.95	**72.73**	kin	52.39	**56.73**	kin	75.28	**78.75**
pcm	50.47	**52.09**	orm	52.28	**61.38**	orm	67.23	**74.11**
por	60.93	**64.81**	pcm	55.39	**59.93**	pcm	64.85	**67.61**
swa	28.26	**61.42**	ptMZ	22.09	**36.80**	som	**55.66**	55.64
tso	35.37	**38.81**	som	48.78	**54.86**	swa	91.51	**91.2**
twi	47.2	**56.00**	swa	30.74	**34.35**	tir	50.2	**55.9**
yor	72.27	**74.63**	tir	57.22	**60.71**	twi	46.89	**48.42**
orm	20.09	**24.28**	vmw	21.18	**22.08**	xho	50.91	**59.17**
tir	22.45	**24.53**	yor	28.65	**39.26**	yor	53.44	**77.9**
Avg.	51.39	**58.21**	**Avg.**	44.25	**51.45**	**Avg.**	65.17	**69.83**

Table 3: Result of baseline (AfroXLMR) and DAPT (AfroXLMR-Social) across the three datasets (AfriSenti, AfriEmo, and AfriHate). During TAPT, the text for the task-adaptive data is without the labels, and the evaluation is cross-tasked among the three target datasets. Reported results are macro-F1.

AfriSocial and AfroXLMR-Social

HCDS HUB OF COMPUTING & DATA SCIENCE

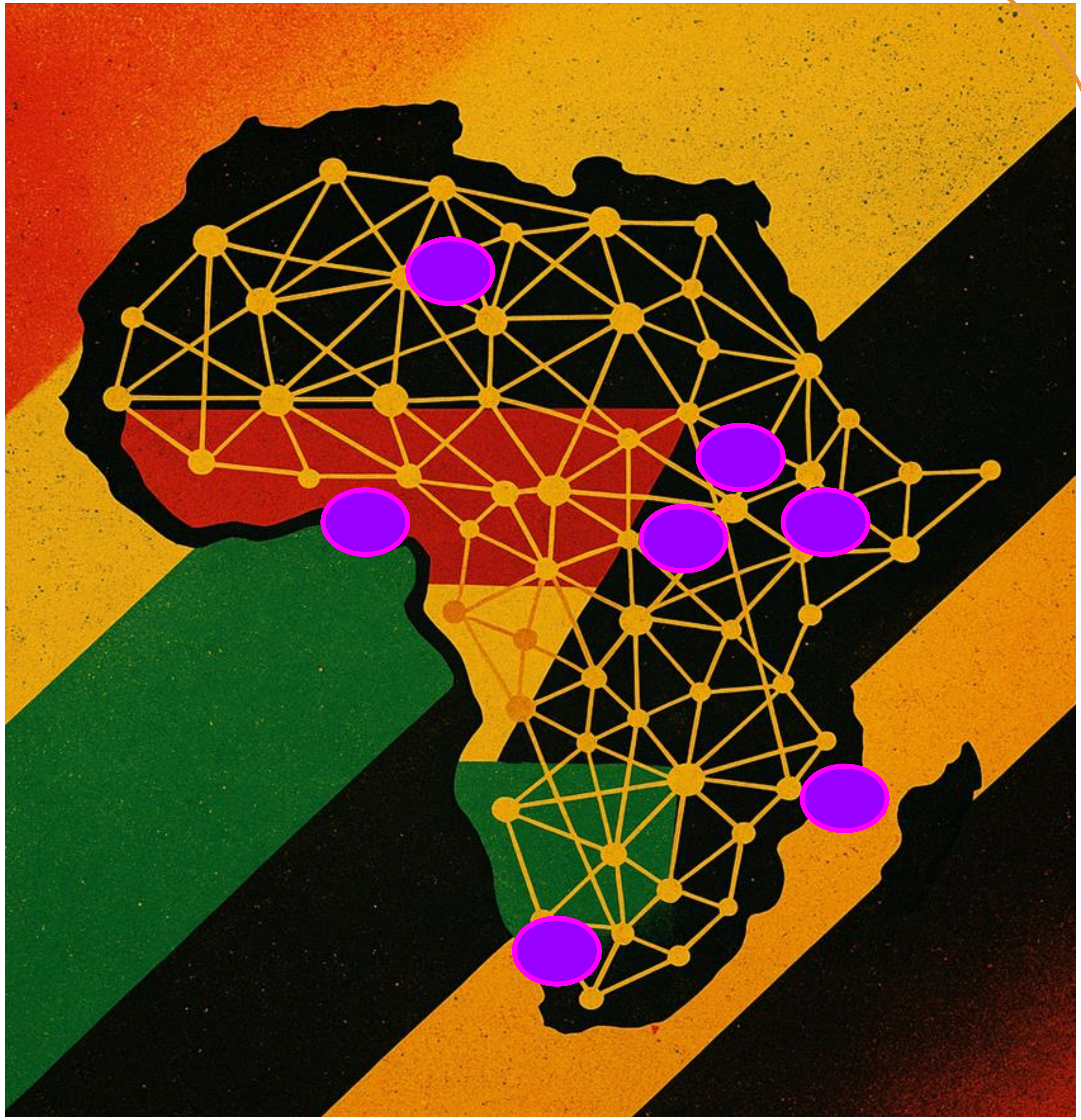
- Introduced **AfriSocial**: new social media/news corpus for 14 African languages.
- Applied domain-adaptive (**DAPT**) and task-adaptive (**TAPT**) pre-training to AfroXLMR for social media NLP tasks.
- Showed DAPT and TAPT consistently improve F1 by **1–30%** for sentiment, emotion, and hate speech classification (19 languages).
- Combined DAPT + TAPT further **boosts performance**.
- **AfroXLMR-Social** outperforms general LLMs on African social media text.

Deep Learning Indaba 2025
Kigali, Rwanda

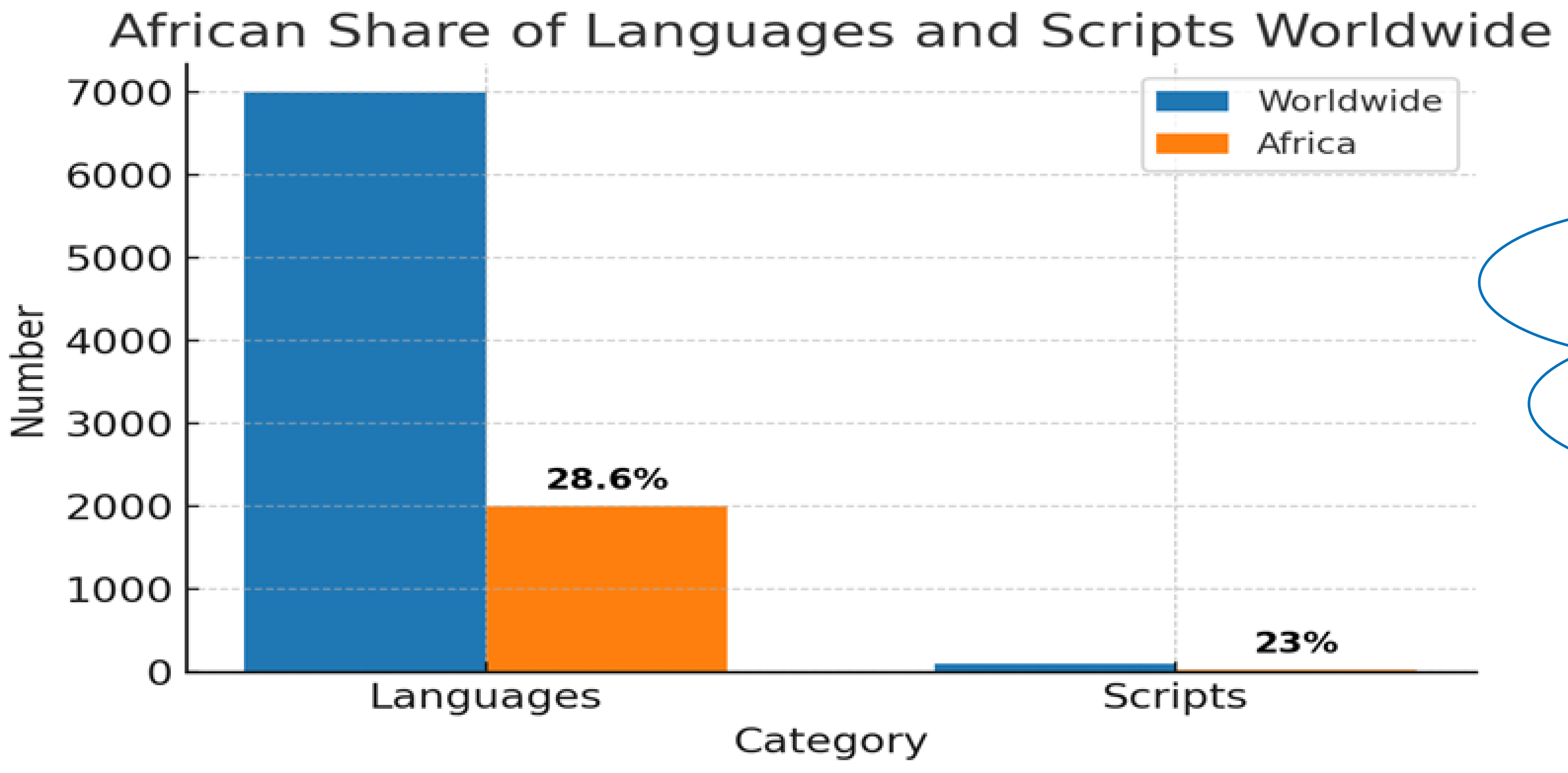
The State of Large Language Models for African Languages

Progress, Challenges & Prospects

Kedir Yassin Hussen[1], Walelign Tewabe Sewunetie[2], Abinew Ali Ayele[3], Sukairaj Hafiz Imam[4], Eyob Nigussie Alemu[5] , Shamsuddeen Hassan Muhammad[4,6] ,
Seid Muhie Yimam[7]



Introduction, Objective, Methodology and Findings



Questions

- ★ Language Coverage
- ★ Script Support
- ★ Dataset Availability
- ★ Task Representation
- ★ Resource Gap
- ★ Prospects

Category	Parameter Range	Examples
LLMs	>7 B	GPT-4, PaLM 2, LLaMA 3
SLMs	500 M–7 B	mBERT, mT5, XLM-R
SSLMs	<500 M	AfriBERTa, AfroLM, EthioLLM

- ★ **Only 41(2%)**
- ★ Only **Latin**, **Arabic** & **Ge'ez** scripts.
- ★ **<10 languages** are getting support frequently
- ★ A total of **18GB** of data is under **23** datasets.
- ★ **Classification** is one of the most extensively explored tasks, while others are often neglected.

Conclusion

HCDS HUB OF COMPUTING & DATA SCIENCE

- African languages are **severely underrepresented** in current language models and script coverage.
- Data scarcity, lack of orthographic standards, and high **computational needs** hinder progress.
- **SSLMs** offer targeted **potential for advancing African NLP** through a staged development roadmap.
- It is very challenging to **quantify** the representation of **African languages** in large language models (LLMs).

BEST PAPER AWARD

DEEP LEARNING INDABA

HCDS HUB OF COMPUTING & DATA SCIENCE

Presented to

Kedir Yassin Hussen, Walelign Tewabe Sewunetie, Abinew Ali Ayele, Sukairaj Hafiz Imam, Eyob Nigussie Alemu, Shamsuddeen Hassan Muhammad and Seid Muhie Yimam

in recognition of the paper

The State of Large Language Models for African Languages: Progress and Challenges

This paper has been selected as the Best Paper presented at the Deep Learning Indaba, Urunana, held in August 2025. This recognition is a testament to the exceptional quality, originality, and profound impact of research.

Ssekiwere Bruno	***Albert Njoroge***
General Chair, Indaba 2025	*Chair, Publications Committee*

Universität Hamburg DER FORSCHUNG | DER LEHRE | DER BILDUNG

Interdisciplinary Collaboration and Co-Creation

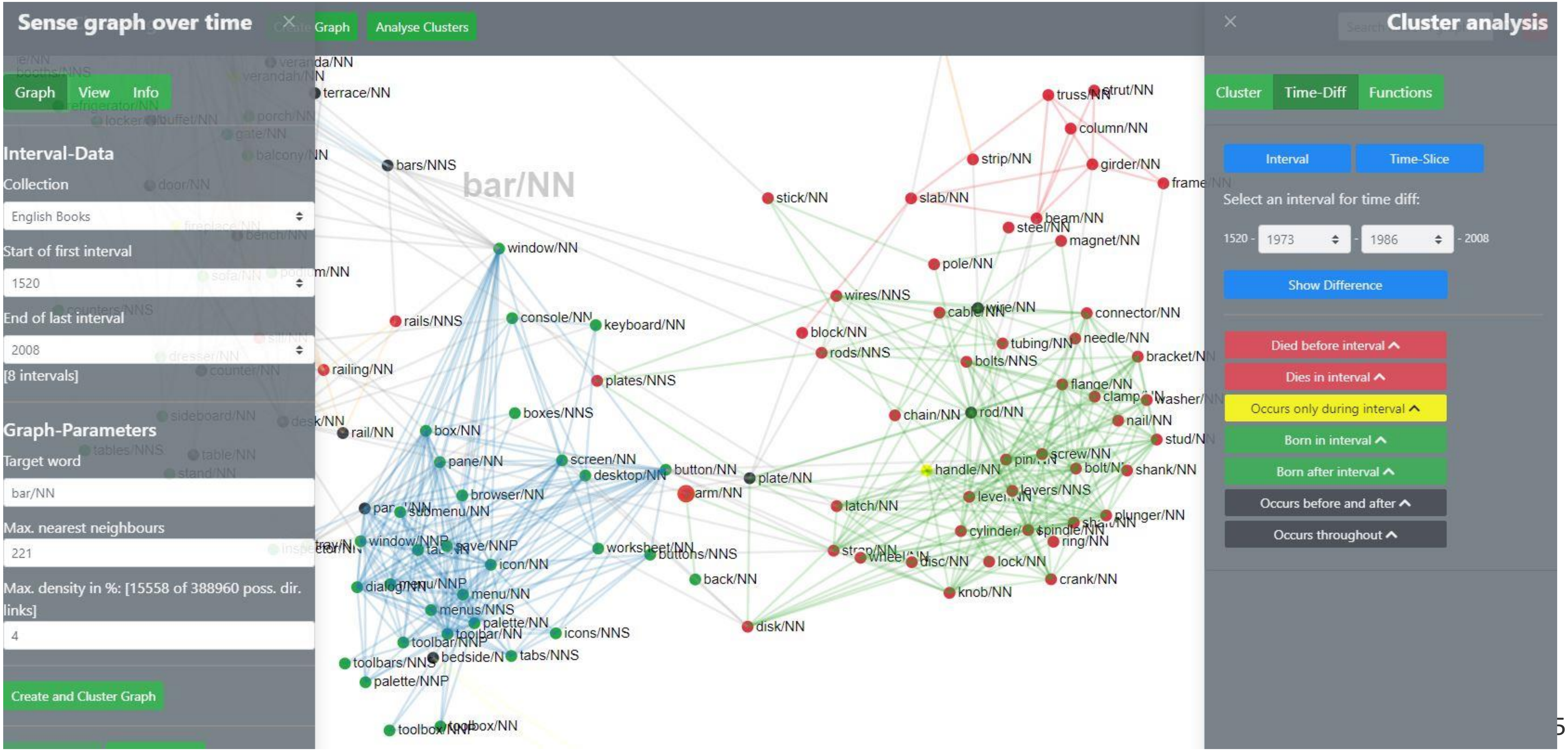
SCoT: Sense Clustering over Time – a tool for analysing lexical change

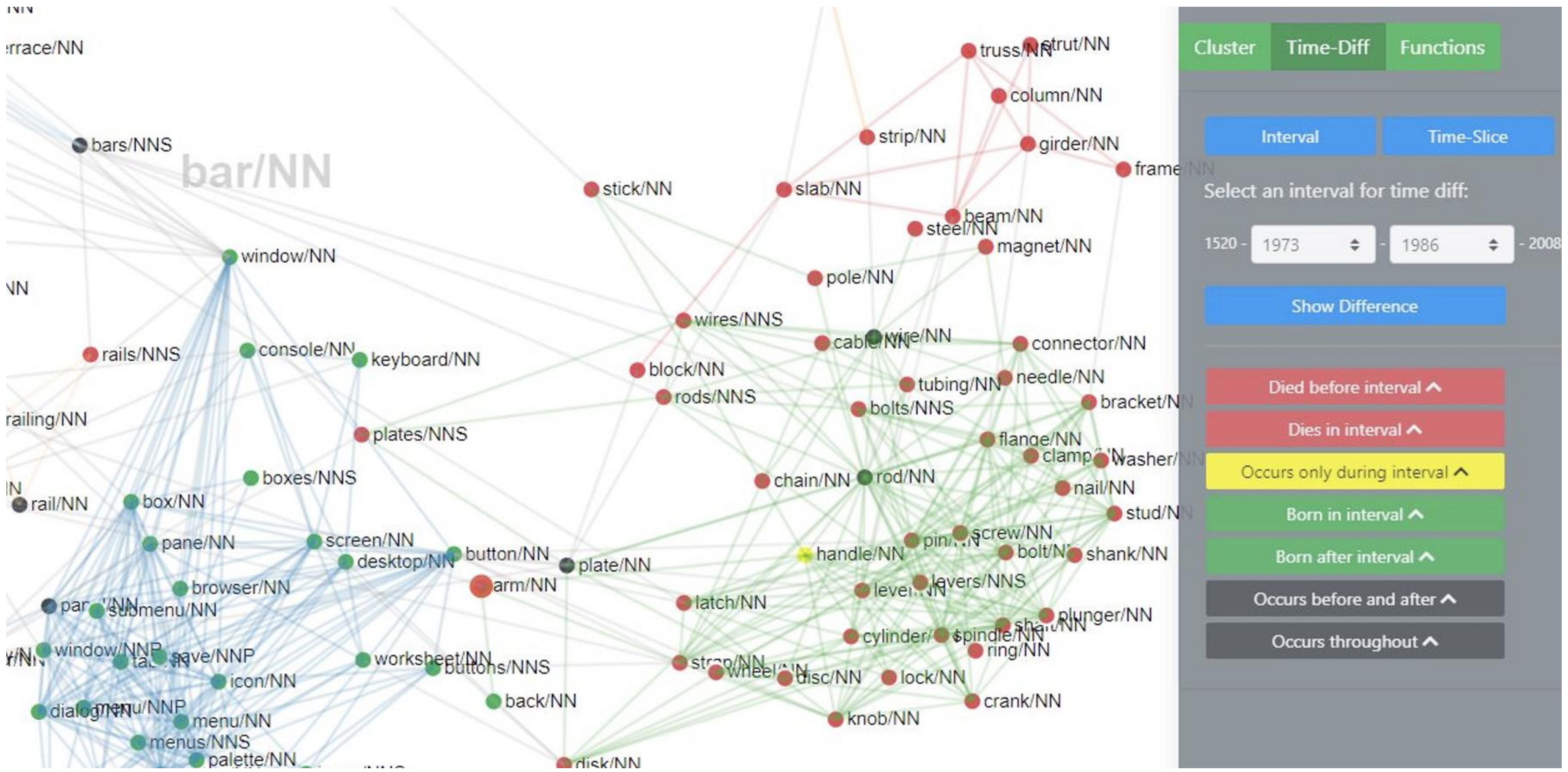
Christian Haase[†], Saba Anwar[†], Seid Muhie Yimam[†], Alexander Friedrich[*], Chris Biemann[†]
[†] Language Technology group, Universität Hamburg, Germany
[*] Institute for Philosophy, TU Darmstadt, Germany
{anwar,yimam,biemann}@informatik.uni-hamburg.de
haase.mail@web.de
friedrich@phil.tu-darmstadt.de

SCOT

- **SCOT:** Analyze how a word's senses (meanings) change across historical periods
- **Approach: JoBimText Framework** (Biemann & Riedl, 2013), **Chinese Whispers Clustering** (Biemann, 2006)
  - Split large corpora into time intervals (e.g., decades)
  - Build a word similarity graph (semantic neighborhood) for each interval
  - Merge graphs into a dynamic, time-aware network, nodes = words, clusters = senses
  - Cluster nodes (words) to identify distinct senses for the target word in each interval
- **Visualization:**
  - Color-coding shows when words/senses appear (green), disappear (red), or persist
  - Sense clusters visually reveal shifts, mergers, and splits in word meaning

Analysis of the sense shifts of ‘bar/NN’ in Google Books (Goldberg and Orwant, 2013) with SCoT





- Analysis of the sense shifts of ‘bar/NN’ in Google Books (Goldberg and Orwant, 2013) with SCoT: the clusters of the neighbourhood graph over time show that the sense “**a rigid piece of metal used as a fastening or obstruction**” [**top right**] loses traction, while the sense ”**computer-menu**” [**bottom left**] gains significance. The coloring is relative to the interval **”1973-1986”. Red** indicates the disappearance of a node before 1986. **Green** indicates the emergence of a node after 1986.

Discourse Analysis Tool Suite

Developed by LT

Developers, contributors, researchers:

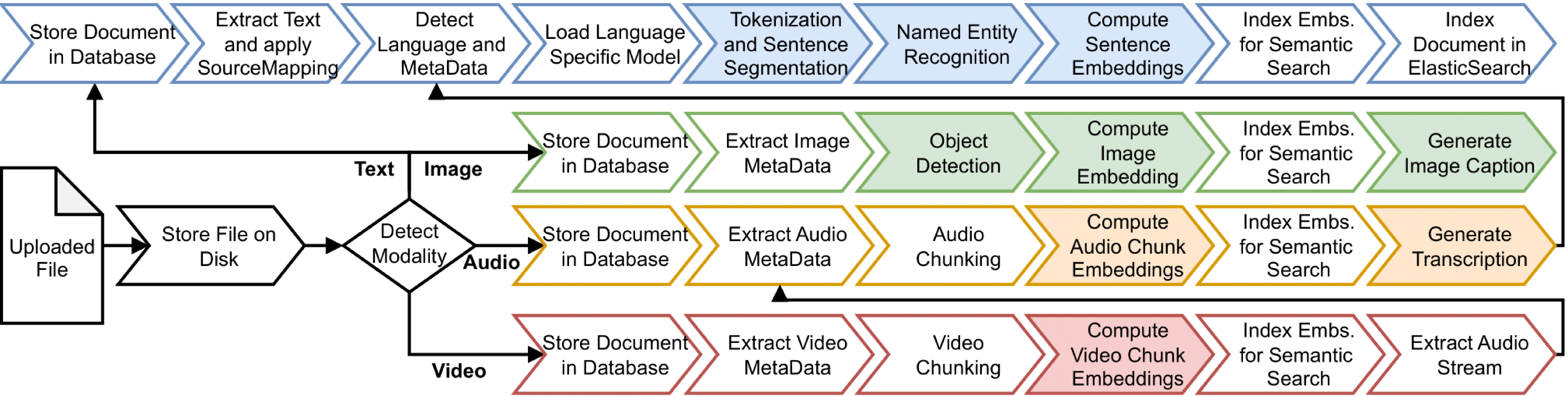
Tim Fischer, Florian Schneider, Fynn Petersen-Frey, Gertraud Koch, Robert Geislinger, Florian Helfer, Anja Silvia Mollah Haque, Isabel Eiser, Martin Semmann, Yannick Walter, Stefan Aykut, Chris Biemann

DATS (Discourse Analysis Tool Suite) – At a Glance

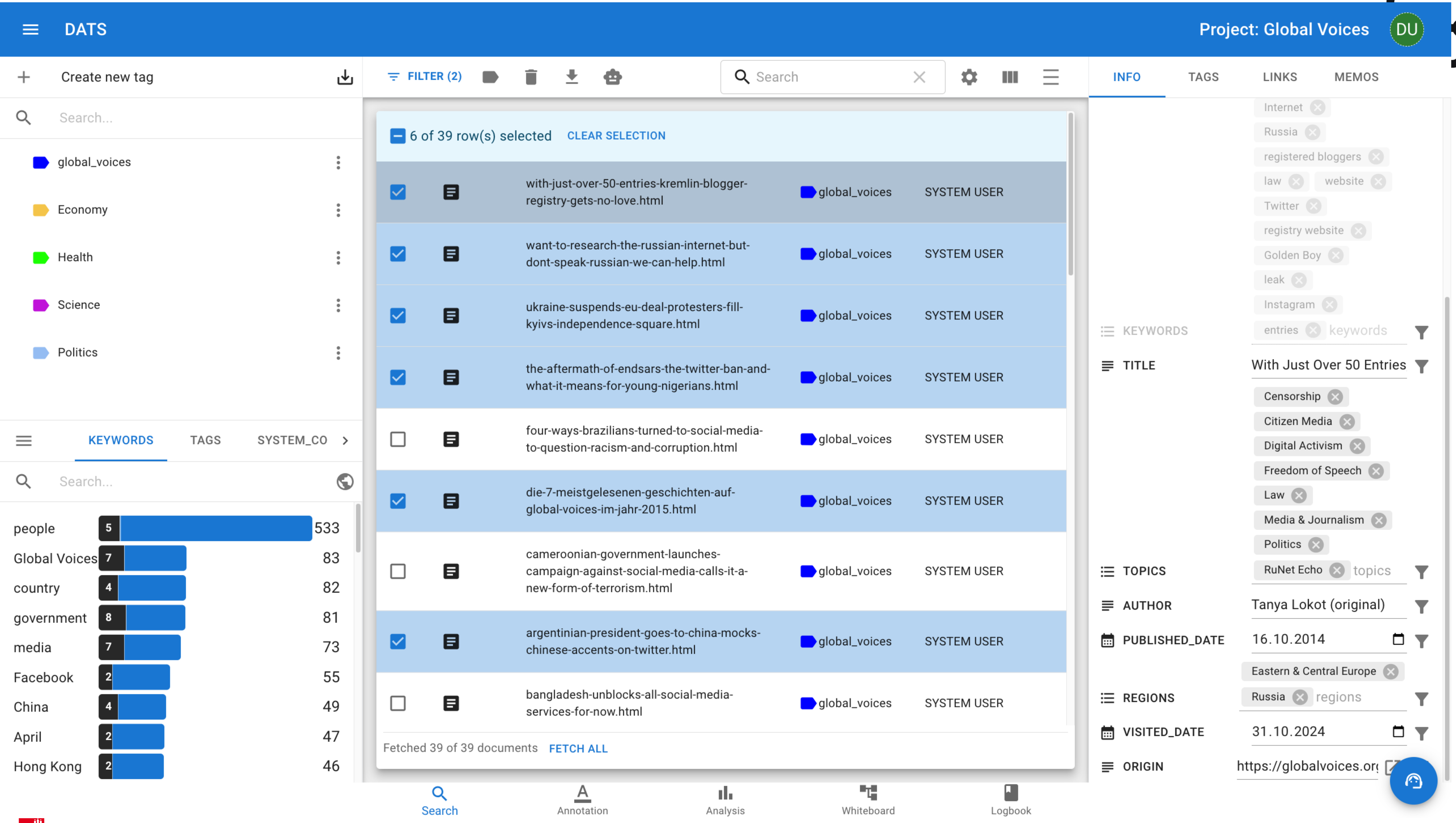
HCDS
HUB OF COMPUTING
& DATA SCIENCE

- Open Source Platform for **qualitative analysis** of large, **multi-modal** data (text, images, etc.)
- **Human-in-the-Loop** AI: Machine learning and NLP tools assist, but users keep control and transparency
- **Streamlined Workflow**: Import, search, annotate (manual/AI-supported), analyze, and visualize data—all in one place
- **Visual & Reflective Tools**: Innovative Whiteboards support visual mapping, memoing, and theory-building
- **Designed for Non-Experts**: No ML experience needed; intuitive UI for Digital Humanities & Social Science researchers
- **Co-created & Expandable**: Developed with scholars, open for collaboration, and easily extended with new features

https://github.com/uhh-lt/dats



Store Document in Database
Extract Text and apply SourceMapping
Detect Language and MetaData
Load Language Specific Model
Tokenization and Sentence Segmentation
Named Entity Recognition
Compute Sentence Embeddings
Index Embs. for Semantic Search
Index Document in ElasticSearch
Uploaded File
Store File on Disk
Detect Modality
Text
Image
Audio
Video
Store Document in Database
Extract Image MetaData
Object Detection
Compute Image Embedding
Index Embs. for Semantic Search
Generate Image Caption
Store Document in Database
Extract Audio MetaData
Audio Chunking
Compute Audio Chunk Embeddings
Index Embs. for Semantic Search
Generate Transcription
Store Document in Database
Extract Video MetaData
Video Chunking
Compute Video Chunk Embeddings
Index Embs. for Semantic Search
Extract Audio Stream



HCDS
HUB OF COMPUTING & DATA SCIENCE

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

LLM Assistant - Document Tagging

Select method — Select tags — Edit prompts — Wait — 5 View results

HCDS HUB OF COMPUTING & DATA SCIENCE

Here are the results! You can find my suggestions in the column *Suggested Tags*.
Now, you decide what to do with them:

- Use your current tags (discarding my suggestions)
- Use my suggested tags (discarding the current tags)
- Merge both your current tags and my suggested tags

Merging strategy: USE CURRENT | USE SUGGESTED | MERGE BOTH

	Document	Current Tags	Suggested Tags	New Tags
☐	Gesundheitswesen: Ver.di ruft zu Warnstreiks auf	*no tags*	Germany Economics	Germany Economics

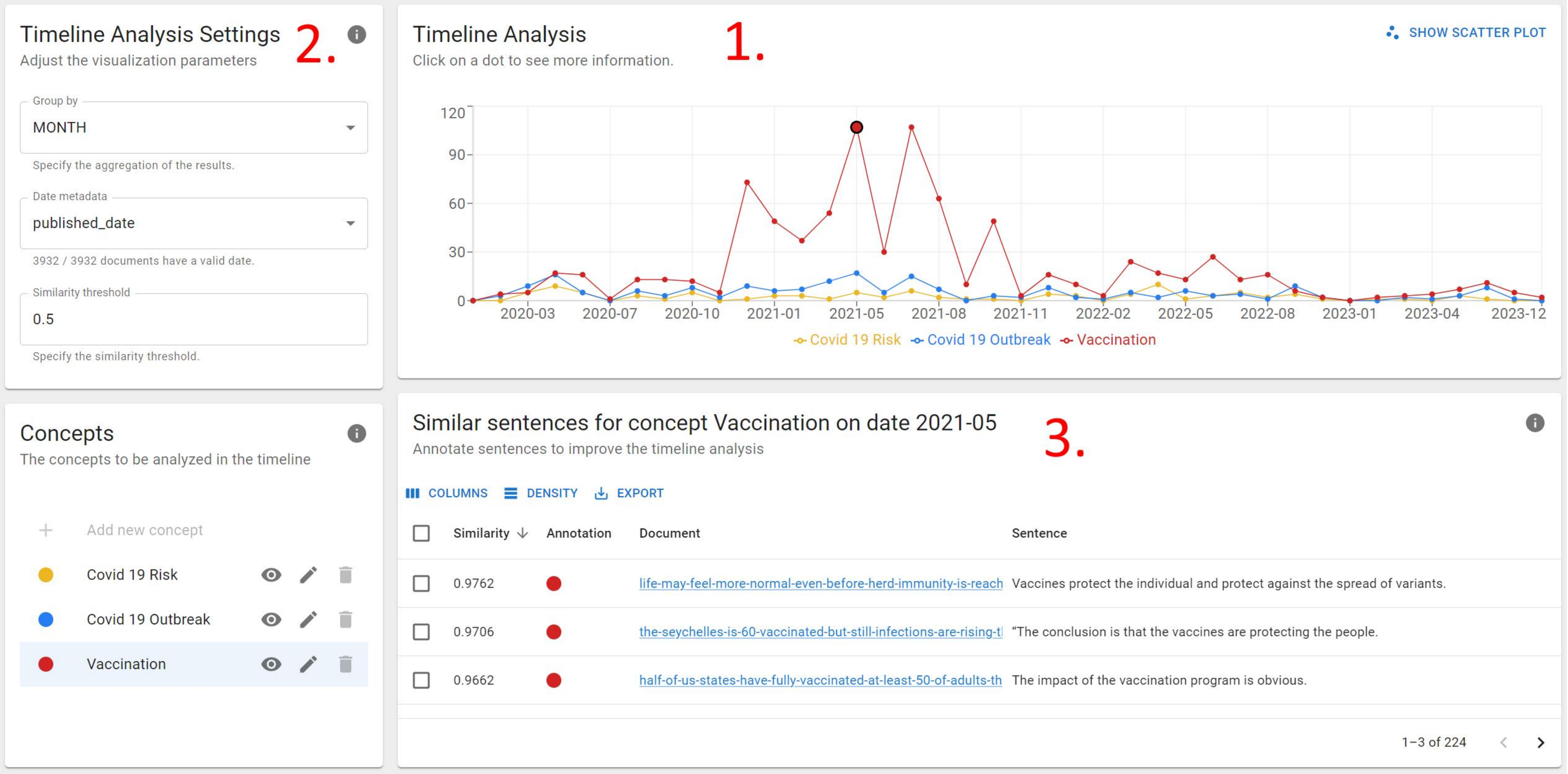
Explanation: The article deals with collective bargaining in the public sector in Germany and the associated warning strikes. It deals with wage increases which is an economic issue.

	Document	Current Tags	Suggested Tags	New Tags
☑	Schutz der Weltmeere: UN einigen sich auf Hochseeabkommen	World	World Science Economics	World Science Economics

1 of 6 row(s) selected CLEAR SELECTION

DISCARD RESULTS & CLOSE | APPLY NEW TAGS

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG


Timeline Analysis Settings
2.
Adjust the visualization parameters
Group by
MONTH
Specify the aggregation of the results.
Date metadata
published_date
3932 / 3932 documents have a valid date.
Similarity threshold
0.5
Specify the similarity threshold.
Timeline Analysis
1.
Click on a dot to see more information.
SHOW SCATTER PLOT
120
90
60
30
0
2020-03
2020-07
2020-10
2021-01
2021-05
2021-08
2021-11
2022-02
2022-05
2022-08
2023-01
2023-04
2023-12
Covid 19 Risk
Covid 19 Outbreak
Vaccination
Concepts
The concepts to be analyzed in the timeline
Add new concept
Covid 19 Risk
Covid 19 Outbreak
Vaccination
Similar sentences for concept Vaccination on date 2021-05
3.
Annotate sentences to improve the timeline analysis
COLUMNS
DENSITY
EXPORT
Similarity
Annotation
Document
Sentence
0.9762
life-may-feel-more-normal-even-before-herd-immunity-is-reach
Vaccines protect the individual and protect against the spread of variants.
0.9706
the-seychelles-is-60-vaccinated-but-still-infections-are-rising-t
“The conclusion is that the vaccines are protecting the people.
0.9662
half-of-us-states-have-fully-vaccinated-at-least-50-of-adults-th
The impact of the vaccination program is obvious.
1–3 of 224

Challenges and Future Directions

State of AI in Africa

Community Driven Participatory research

Compared to other regions where the AI ecosystem is shaped by Universities, big corporations or strong policies and regulation frameworks:

Africa's AI ecosystem is dominated by **grassroots movements**, such as 'Deep Learning Indaba' and 'Data Science Africa'.

AI and the Future of Work in Africa White Paper O'Neill, J., et al. (2024, June).

Blossoming of Local Communities

HCDS HUB OF COMPUTING & DATA SCIENCE

Addressing the challenges of NLP for African languages through participatory approach

DEEP LEARNING INDABA

DSN Data Science Nigeria

Ou Ethio NLP

NLP Corpus and Dataset Creation

Language Model Building

Research & Collaboration

Assist Education Quality

Academy to Industry Linkage

Marketplace for Professionals

Major Challenges in Modeling Low-Resource Humanities Data

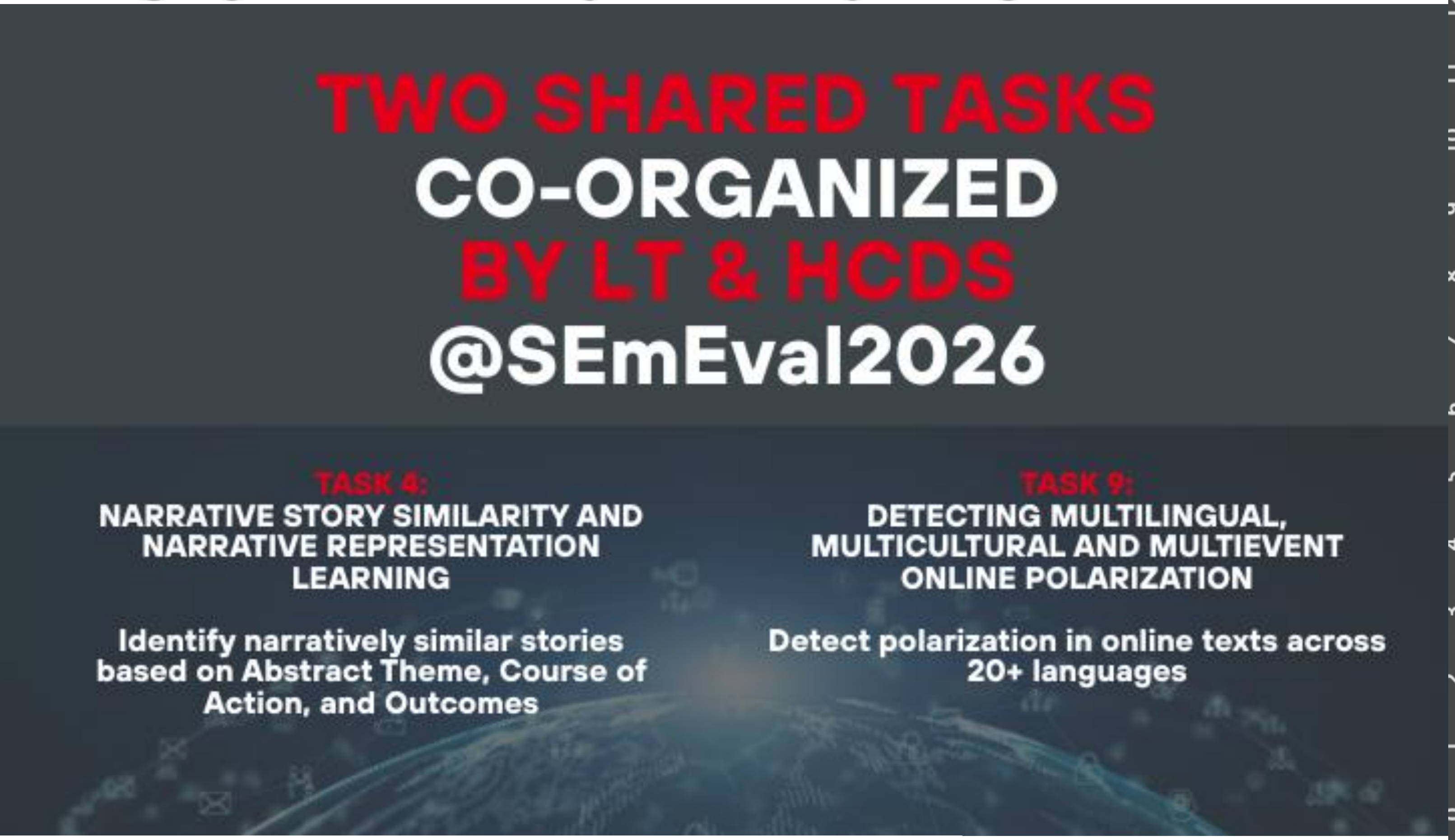
- Data Scarcity & Imbalance
  - Most datasets are **small**, with many languages underrepresented.
- Linguistic & Cultural Diversity
  - Languages with **complex scripts**, tonal features, dialects, and code-switching complicate modeling.
- Limited Resources & Infrastructure
  - **Funding**, **infrastructure**, and open data access remain major barriers, particularly across African languages.
- Evaluation & Standards
  - Predominance of surface metrics (e.g., BLEU, WER) limits understanding of **linguistic** and **cultural** nuances.

Key Takeaways for Low-Resource Humanities AI

- **Interdisciplinary** & Community-Driven
  - Collaboration with **linguists**, cultural **experts**, and local **communities** is essential.
- Language & Culture-Aware Models
  - Technologies should **preserve dialectal**, **stylistic**, and cultural diversity.
- Data & Resource Development
  - Prioritize **open-access**, high-quality, and multi-dimensional datasets.
- Models:
  - SLM/SSLM are still popular, **LLMs** are limited (especially for generation tasks)
- Future Focus
  - Invest in **scalable infrastructure**, **ethical frameworks**, and context-sensitive evaluation methods.

📣 Announcements 🔔

SemEval-2026



KONVENS 2026

HCDS HUB OF COMPUTING & DATA SCIENCE

Website coming soon

KONVENS 2026

CONTEXT MATTERS: NLP BEYOND TEXT

konvens26.hcds.uni-hamburg.de

People

Prof. Dr. Chris Biemann
Language Technology

HCDS HUB OF COMPUTING & DATA SCIENCE

Prof. Dr. Anne Lauscher
Data Science

Prof. Dr. Heike Zinsmeister
Corpus Linguistics

HCDS
HUB OF COMPUTING
& DATA SCIENCE

fin