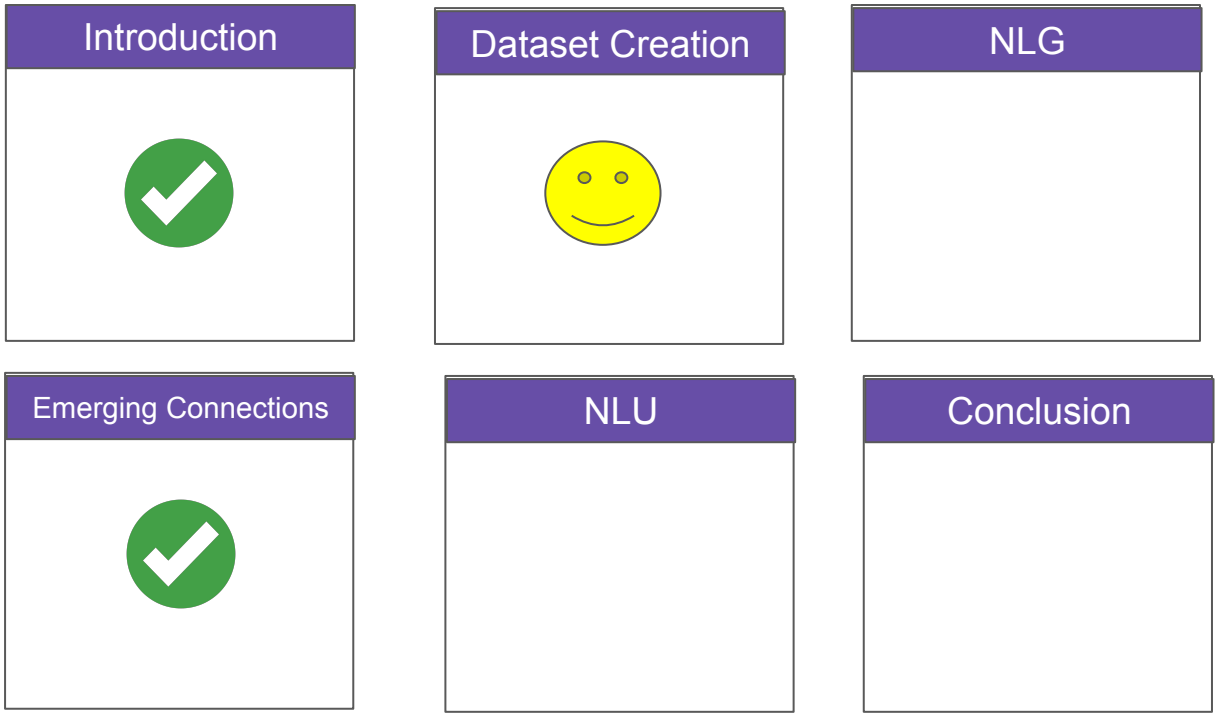
Connecting Ideas in ‘Lower-Resource’ Scenarios: NLP for National Varieties, Creoles and Other Low-resource Scenarios

Aditya Joshi, Diptesh Kanojia, Heather Lent, Hour Kaing, Haiyue Song

Tutorial Agenda



Module 3: Common Ideas in Dataset Creation (40 minutes)

– Identifying Data Sources

– Computationally Assisted Annotation Tools

– Existing Multilingual and Code-Mixing Datasets

– Case studies:

- Synthetic Data
- Translation

– Q&A & Discussion (10 mins)

Identifying Data Sources

Considerations for Identifying Data Sources

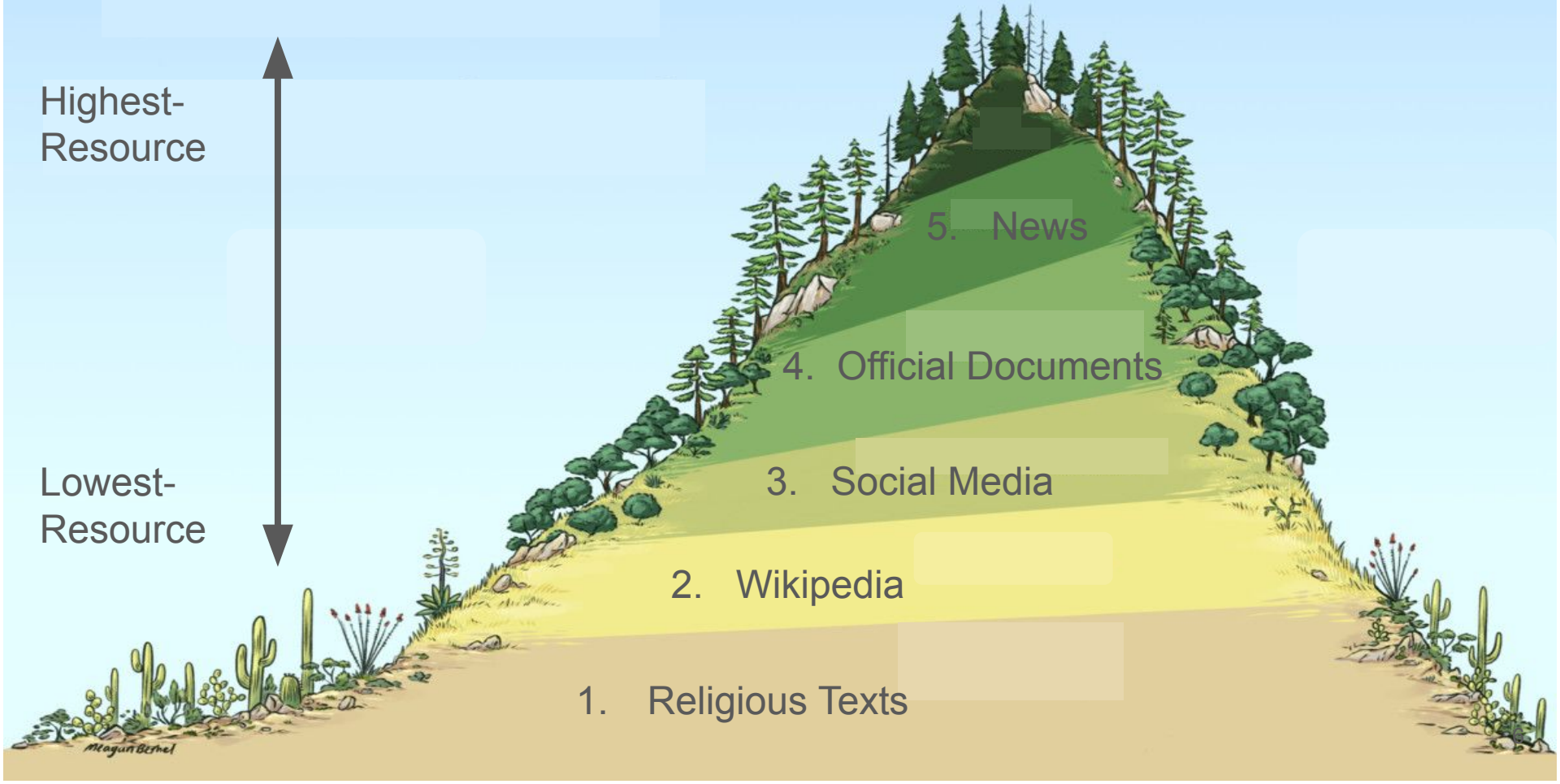
Given a language or language variety, ask yourself …

- Is it situated in a **monolingual** or **multilingual** context?
- Do speakers generally **speak another** (prestigious) **language**?
- Does it have **official status**?
- Is it generally **accepted** or **marginalized**?

Emirati Arabic	**Haitian Creole**	**Singlish**	**Hinglish**
Multi	Mono	Multi	Bilingual*
Yes	No	Yes	Maybe/No
Yes	Yes	No	No
😁	😕	😐	😐

Considerations from [Lent et al. 2022]

The Landscape of Data



Religious Texts

- “Creating a Massively Parallel Bible Corpus” (Mayer and Cysouw, 2014)
  - **900** translations in more than **830** language varieties
- “JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages” (Agić and Vulić, 2019)
  - **300** languages with around 100 thousand parallel sentences per language pair on average
- The Johns Hopkins University Bible Corpus: **1600+** Tongues for Typological Exploration (McCarthy et al., 2020)
  - **4000** translations
- “JWSign: A Highly Multilingual Corpus of Bible Translations for more Diversity in Sign Language” (Gueuwou et al., 2023)

⚠️ Copyright → Not Public; ⚠️ Possible Data Contamination

Wikipedia

- **339** languages of varying sizes
  - ⚠️ Data dumps not available for all
  - ⚠️ Typically standardized forms

- ⚠️ Quality: Your Mileage May Vary
  - “Quality at a Glance” by Kreutzer et al. 2022
  - “How Good is Your Wikipedia” by Tatariya et al. 2024

- ⚠️ Direct & Indirect data contamination: models (mBERT), langID (CC4)

- ⚠️ Check if machine generated
  - https://stats.wikimedia.org/EN/BotActivityMatrixCreates.htm

Wikipedia

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	All languages		Russian			Spanish				Polish			Czech			Korean
		English			Japanese			Italian			Portuguese			Swedish		
			German			French			Chinese			Dutch			Arabic	
	Σ	**en**	**de**	**ru**	**ja**	**fr**	**es**	**it**	**zh**	**pl**	**pt**	**nl**	**cs**	**sv**	**ar**	**ko**
Σ total articles created	49.3 M	5.8 M	2.3 M	1.5 M	1.1 M	2.1 M	1.5 M	1.5 M	1.0 M	1.3 M	1.0 M	2.0 M	419 k	3.8 M	653 k	439 k
Σ manually created articles	33.1 M	5.6 M	2.3 M	1.4 M	1.1 M	2.0 M	1.5 M	1.4 M	877 k	1.2 M	861 k	907 k	416 k	704 k	455 k	434 k
Σ articles created by bots	**16.2 M**	**150 k**	**1.2 k**	**149 k**	**140**	**79 k**	**2.2 k**	**110 k**	**161 k**	**164 k**	**153 k**	**1.0 M**	**3.0 k**	**3.1 M**	**198 k**	**4.5 k**
Share of articles created by bots	**32.9%**	**3%**	**0%**	**10%**	**0%**	**4%**	**0%**	**7%**	**16%**	**12%**	**15%**	**54%**	**1%**	**81%**	**30%**	**1%**
Lsjbot	**9.5 M**		1											3.0 M		
Cheers!-bot	**561 k**															
Dcirovicbot	**539 k**															
Joopwikibot	**524 k**											524 k				
LymaBot	**213 k**											213 k				

Wikipedia

56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
	Uzbek		Malayalam			Latin			Waray-Waray				Welsh			Occitan			Kyrghyz			Bret n		Al
s		Nynorsk			Belarusian			Telugu			Ast urian		Egyptian Arabic			Mongolian			Luxem bourgish				Tatar	
acedon ian			Marathi			Cantonese			Cebuano			Icelandic			Scots			Kanna da			Swahili			Ar gones
mk	uz	nn	mr	ml	be	zh_yue	la	te	ceb	war	ast	is	cy	arz	sco	oc	mn	kn	ky	lb	sw	br	tt	an
98 k	130 k	142 k	53 k	62 k	160 k	70 k	130 k	70 k	5.4 M	1.3 M	100 k	46 k	102 k	20 k	54 k	85 k	21 k	24 k	80 k	55 k	48 k	66 k	83 k	34 k
80 k	12 k	140 k	53 k	62 k	128 k	70 k	130 k	41 k	22 k	121 k	26 k	46 k	84 k	20 k	54 k	48 k	21 k	24 k	35 k	54 k	48 k	66 k	32 k	30 k
18 k	**118 k**	**1.8 k**	**128**	**3**	**32 k**	**3**	**4**	**29 k**	**5.4 M**	**1.1 M**	**74 k**	**12**	**18 k**	**-**	**5**	**37 k**	**273**		**45 k**	**1.6 k**	**96**		**51 k**	**4.2 k**
19%	**91%**	**1%**	**0%**	**0%**	**20%**	**0%**	**0%**	**42%**	**100%**	**90%**	**74%**	**0%**	**18%**	**-**	**0%**	**44%**	**1%**		**56%**	**3%**	**0%**	**0%**	**62%**	**12%**
									5.3 M	1.1 M														

⚠️ Some Wikipedias are 50 – 100% generated.

Rise of LLM-Generated Datasets

- **Anecdote**: Increasing submissions with NMT-made datasets…
- "**Translationese**": The language of NMT is known to be less rich (*e.g.*, lexical variety) than human-generated text.
  - Leads to inflated performance.
  - "Multi-perspective Alignment for Increasing Naturalness in Neural Machine Translation" (Lai et al. 2024).
- Open Questions:
  - **Quality vs Quantity?**
  - Bottom Line: Usability for End Users
  - "Changing the World by Changing the Data" (Rogers, 2021)

Social Media

Facebook & Instagram (Meta Content Library and API)

- “Creating a Lexicon of Bavarian Dialect by Means of Facebook Language Data and Crowdsourcing” (Burghardt et al., 2019)
- “HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection” (Vargas et al., 2022)

TikTok (⚠️ Research API for **“qualifying”** researchers in US? and Europe)

- “DanTok: Domain Beats Language for Danish Social Media POS Tagging” (Hansen et al. 2023)

Twitter/X (⚠️ No more free API access for academics)

- “Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter” (Okky & Budi, 2019)
- “MMT: A Multilingual and Multi-Topic Indian Social Media Dataset” (Dalal et al. 2023)
- “Evaluating ChatGPT and Bard AI on Arabic Sentiment Analysis” (Al-Thubaity et al., 2023)

Official Documents

- Official documents can assist in training LLMs (*e.g.*, Kummervold et al., 2022)
- Some countries have API's for accessing docs (*e.g.*, Koeva et al., 2020)
- Formatting (e.g., 40 C.F.R. § 1508.1(m)) can help with automatic alignment across languages (*e.g.,* Ploeger et al., 2025 *to appear at NoDaLiDa*)
  - ⚠️ Domain
- Massively multilingual official documents via UN (*e.g.,* EuroParl or UNDHR)
  - ⚠️ Small data sizes for some languages
  - ⚠️ Standardized version of the language
- ⚠️ Country =/= Language

News

Provider	URL	Language	License
Selkosanomat	https://selkosanomat.fi/	fi	CC BY-NC-ND 4.0
Journal Essentiel	https://journalessentiel.be/	fr-BE	CC BY-SA 4.0
Informazione Facile	https://informazionefacile.it/	it-IT	CC BY-SA 4.0
Lätta Bladet	https://ll-bladet.fi/	sv-SE	CC BY-NC-ND 4.0
The Times in Plain English	https://www.thetimesinplainenglish.com/	en-US	*"may be reproduced and distributed by all"*
Infoeasy	https://infoeasy-news.ch/	de-CH	CC BY-NC-ND 4.0

Table 2: List of providers of the news articles that constitute the corpus.

"A Multilingual Simplified Language News Corpus" (Hauser et al., 2022)

- WMT'19 (Barrault et al., 2019)
- AI4Bharat Indic NLP Corpus (Kunchukuttan et al., 2020)
- MasakhaNEWS (Adelani et al., 2023)
- NSina: A News Corpus for Sinhala (Hettiarachchi et al., 2024)

* Dictionaries, Lexicons & Other Directions

- Online dictionaries to create dialectal lexicons (e.g., French and Algerian French; Azouaou et al (2017))
- Location-based filtering: Use locations to determine dialects; Often needs additional steps (See discussion in Gouette et al, 2016)
- Keyword-based filtering: Use a seed set of terms to crawl texts (Singaporean English; Wang et al (2017))
- Annotation: Native speakers of language varieties

[1]Faical Azouaou and Imane Guellil. 2017.Alg/fr: A step by step construction of a lexicon between algerian dialect and french. @PACLIC, Vol. 31.
[2]Hongmin Wang, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu. 2017.Universal Dependencies Parsing for Colloquial Singaporean English. @ACL.
[3]Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016.Discriminating similar languages: Evaluations and explorations. @LREC.

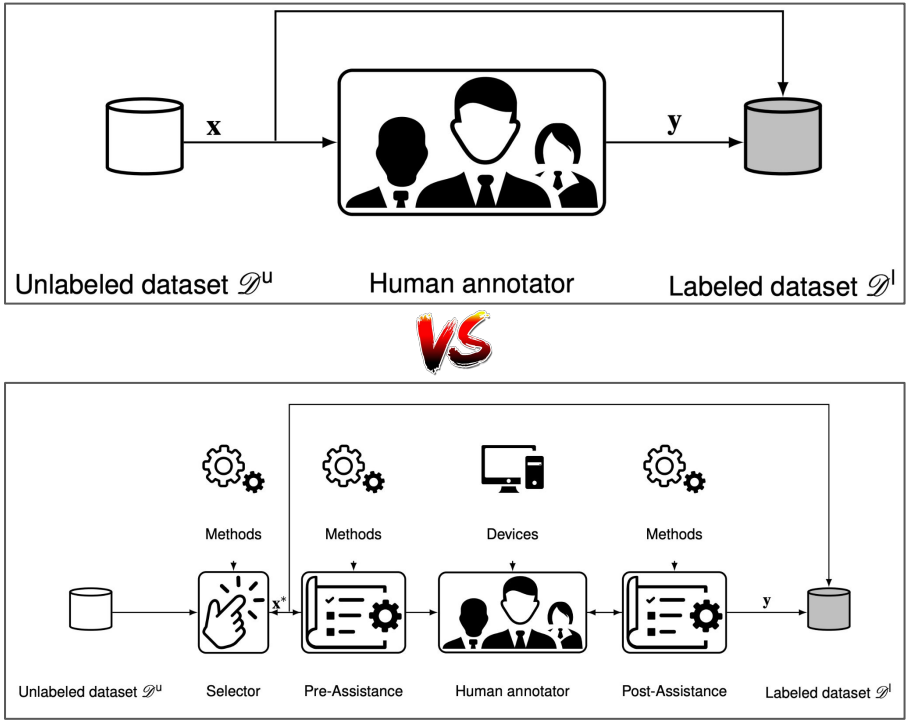
Summary

Resource	Pros	Cons
*		
News	High Quality	Copyright, Standard Dialect
Official Docs.	High Quality	Domain, Standard Dialect
Social Media	Natural	Toxicity & Bias, Privacy, Access
Wikipedia	General Domain, KB	Quality, “Translationese”
Bible	Massively Parallel	Domain, Non-native translations

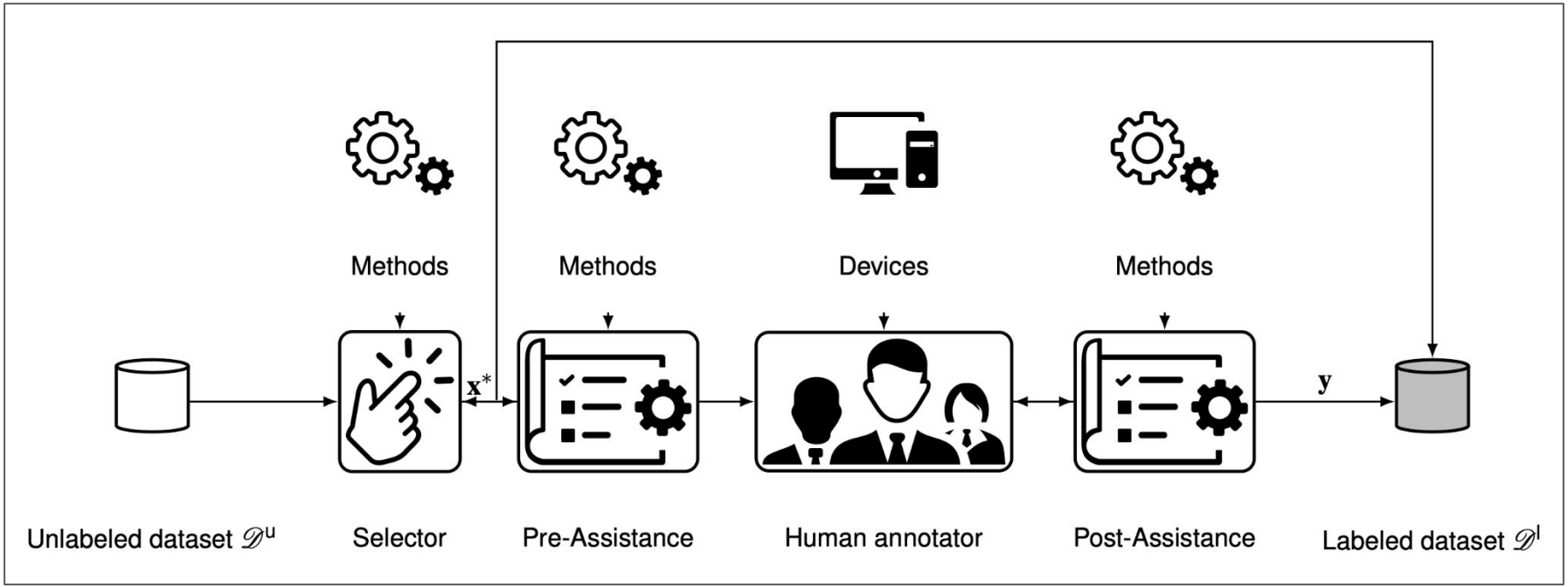
Annotation Tools

Data Annotation | Starting from scratch

- Offline *vs.* Online tools
  - Annotation tools can improve productivity with relevant support, and a user friendly interface.
- Annotation Challenges
  - Subjectivity-- Descriptive *vs.* Prescriptive annotation paradigms (Rottger et al., 2022)
  - Guidelines-- Common set of guidelines + annotator-specific iterative amendments.
  - Workflow-- Naïve *vs.* Assisted annotation workflows (Schilling et al., 2021)
  - Validation-- Regular discussions mitigate consistency issues, address biases, subjective judgements, improves guidelines.
  - Domain Expertise-- Critical for domain-specific data from healthcare, legal, financial, and so on.



Assisted Annotation Workflow



LabelStudio

Label every data type.

GENAI IMAGES AUDIO **TEXT** TIME SERIES MULTI-DOMAIN VIDEO

Sequence Classification Supports taxonomies of up to 10000 classes

Token Classification Extract relevant information into categories

Person[1] Fact[2] Date[3] Time[4] Ordinal[5] Product[6] Language[7] Location[8]

Opossums[Person] are usually solitary[Fact] and nomadic, staying in one area as long as food and water are easily available. Some families will group together in ready-made burrows or even under houses. Though they will temporarily occupy abandoned burrows, they do not dig or put much[Ordinal] effort into building their own. As nocturnal animals, they favor dark, secure areas. These[Date] areas may be below ground or above. When threatened or harmed, they will "play possum", mimicking the appearance and smell of a sick or dead animal.[Product] This physiological response is involuntary (like fainting), rather than a conscious act. In the case of baby opossums, however, the brain does not always react this way at the appropriate moment, and therefore they often fail to "play dead" when threatened.[Language] When "playing possum", the animal's lips are drawn back, the teeth are bared, saliva foams around the mouth, the eyes, close or half-close, and a foul-smelling fluid is secreted from the anal glands. Their stiff, curled form can be prodded, turned over, and even carried away without reaction. The animal will typically regain consciousness after a period of between 40 minutes and 4 hours, a process[Time] which begins[Location] with a slight twitching of the ears.

Quick Start

PIP BREW GIT DOCKER

```
# Install the package
# into python virtual environment
pip install -U label-studio

# Launch it!
label-studio
```

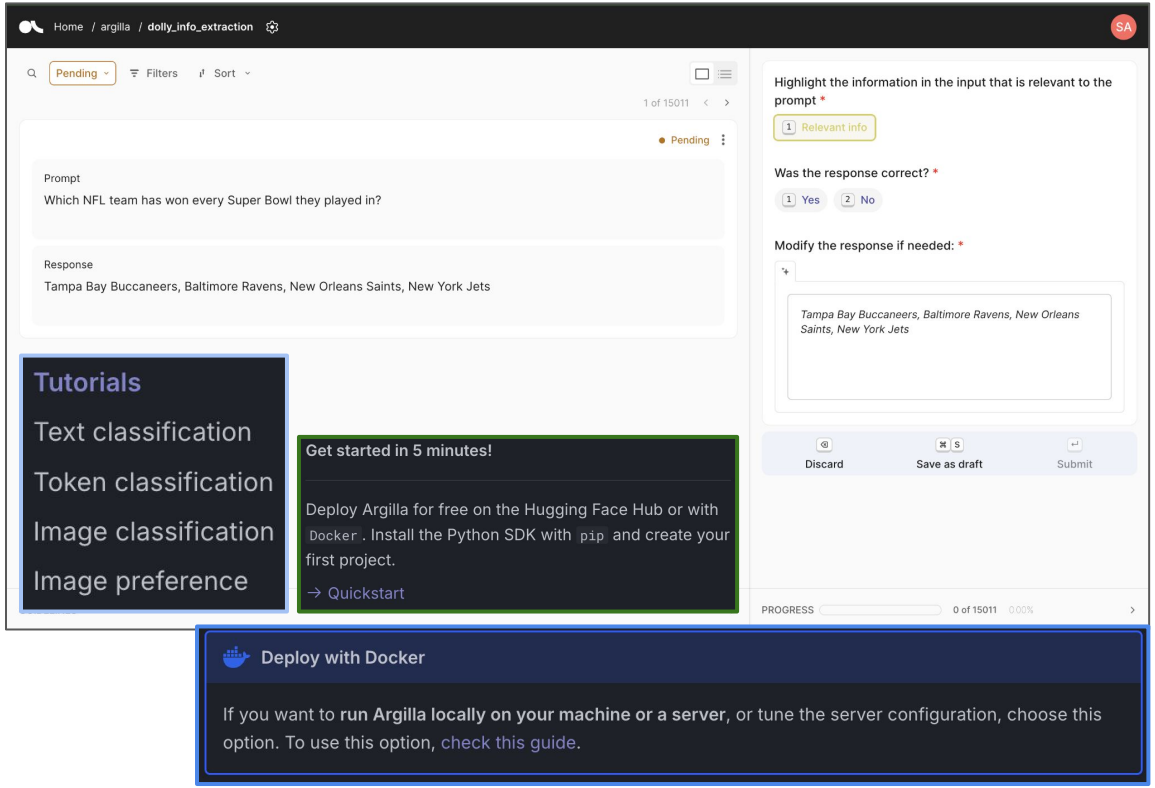
Quick Start

PIP BREW GIT DOCKER

```
# Run latest Docker version
docker run -it -p 8080:8080 -v
`pwd`/mydata:/label-studio/data
heartexlabs/label-studio:latest
```

Argilla

- Supports Active Learning
  - Utilize fine-tuned language models to obtain predictions on data.
  - Model annotation quality improves as more data is annotated.
  - Faster Classical ML approaches
- Python-based backend
  - Data view in *pandas*
  - Leverage EDA libraries
- HuggingFace Integration
  - Deployed to spaces; access to models



distilabel

Demo Video: https://www.youtube.com/watch?v=FIJ6hrBB2bU

Shoonya is an open source platform to annotate and label data for Indic languages.

Offers support for multiple data types (monolingual / parallel datasets) and labeling tasks (NLU, NLG, OCR, ASR, TTS, *etc.*)

Workplace Management

Hierarchical way to manage language work into different organizations, workspaces, and projects.

NMT support

Populating automatic translations from IndicTrans model. Currently supporting 12 Indic languages.

Transliteration Support

Simplified input entry in Roman character with transliteration from IndicXlit models supporting 20+ languages.

Maker-Checker Flow

Multiple ways to evaluate the quality of translated data with automated maker-checker flows.

Context View

View paragraph level context when translating an individual sentence.

Cross-lingual Support

For low-resource language, Shoonya supports showing annotators translations in other languages.

Demo Video: https://www.youtube.com/watch?v=N4PBSB2fQto

Datasets

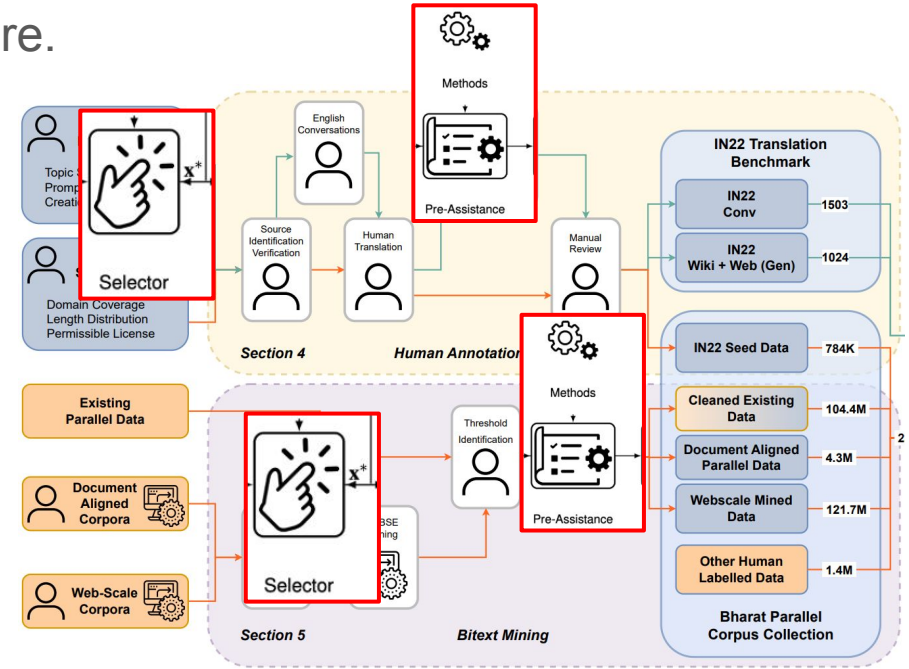
Multilingual,
Multi-Dialect
Code-Mixed
&
Creoles

Multilingual Data

- Massively parallel Bible, UNPC, Wikipedia, Bharat Parallel Corpus Collection (BPCC) [Gala et. al. (2023)], and many more.

Case Study - NLG - BPCC

- BPCC was curated using Shoonya.
- Human Translation
- Weak Supervision
  - Toxicity / LID filtering.
  - Comparable to Parallel Corpora
    - LaBSE based document alignment
- Manual Review

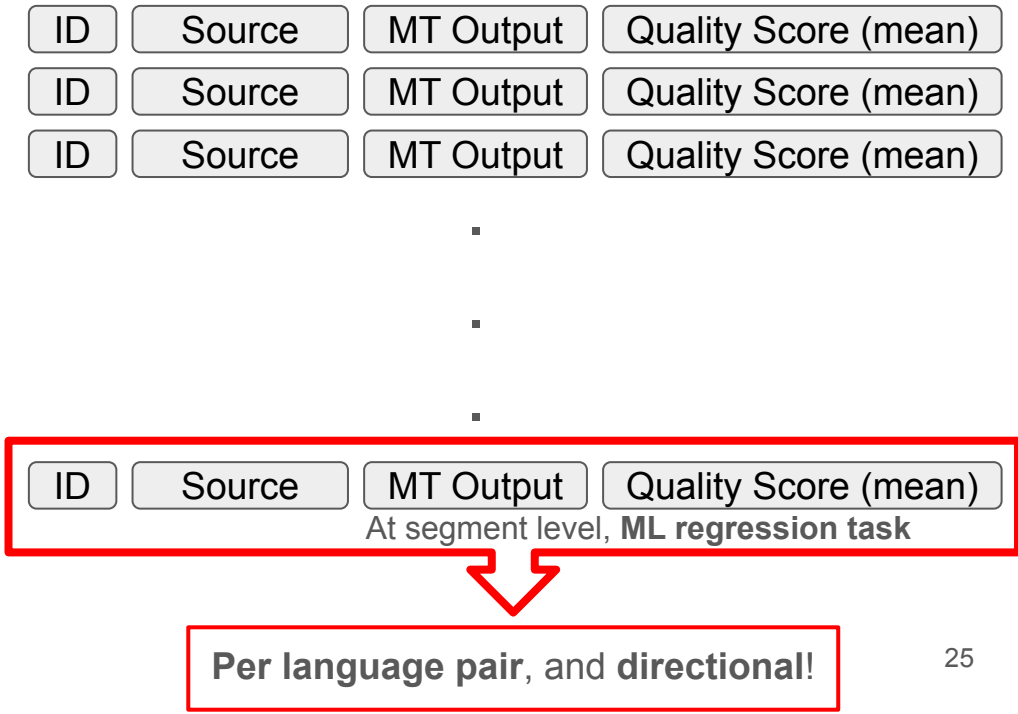


Multilingual Data

Existing large scale multilingual benchmarks
XTREME (Hu et al., 2020) / XTREME-R (Ruder et al., 2021),
XGLUE (Liang et al., 2020), MLQE-PE (Fomicheva et al., 2020, 2022)

Case Study: Quality Estimation (QE) for Machine Translation

- Challenging cross-lingual language understanding task for computational models with following factors to consider:
  - Meaning transfer from source (semantic) and *adequacy*.
  - Structural validity (syntactic) and *fluency*.
  - Domain-specificity / Terminology
  - Transliteration
  - Cross-lingual transfer can contain *cultural nuances*.
  - Idiomatic Translations



Multilingual Data

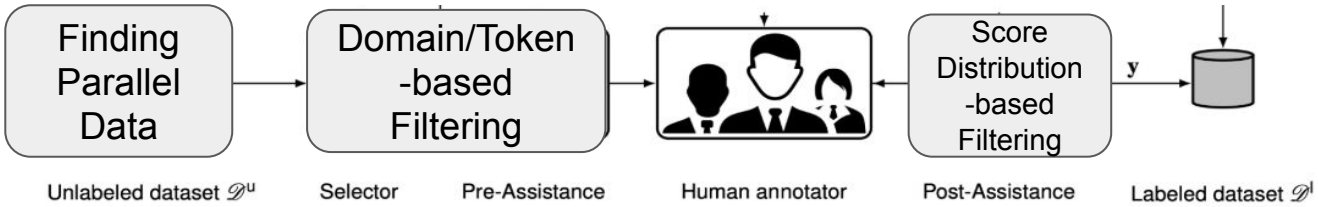
Case Study: Quality Estimation (QE) for Machine Translation [English to Low-resource Indic Languages]

- Since 2021, Surrey NLP[1] has contributed datasets for:
  - English ***to*** Marathi / Hindi / Gujarati / Tamil / Telugu
- Challenges
  - Annotation Guidelines (generic *vs. error dependent*)
  - **M**ultidimensional **Q**uality **M**etrics (cost/expertise)
  - Domain coverage *vs.* annotation expertise
  - Token-based filtering helps maintain annotation cost
  - Extra instances for balancing label distribution
  - **Under construction:** Large-scale X-NLU benchmark

[1] https://huggingface.co/surrey-nlp

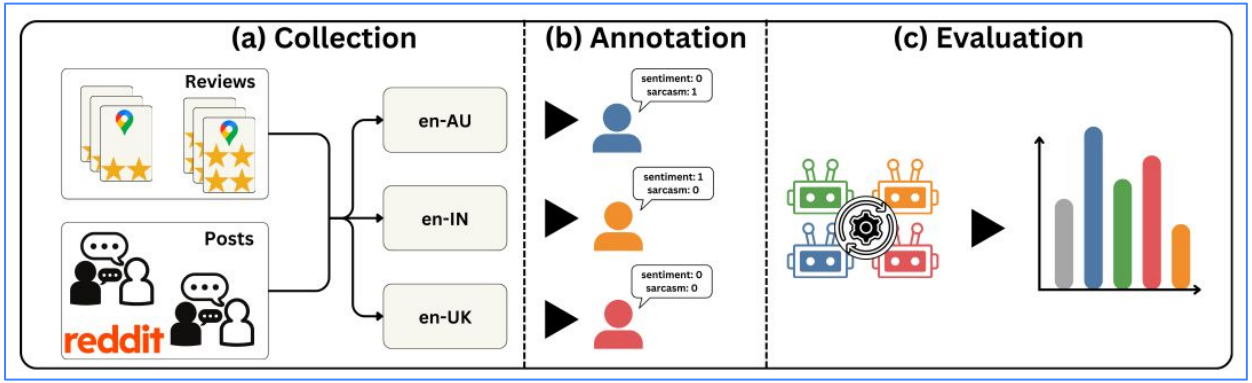
ID	Source	MT Output	Quality Score (mean)

Per language pair, and **directional**!



Multi-Dialect Data

Case Study - NLU
BESSTIE [Srirag et al., 2024]



- Benchmark dialectal dataset for varieties of English with ***both sentiment and sarcasm labels***.
- ***Multi-domain coverage*** as data from Google Places Reviews with weak labels (rating), along with Reddit data from subreddits with a local context.
- **Human annotation** for en-AU, en-UK, en-IN with the help of native speakers.

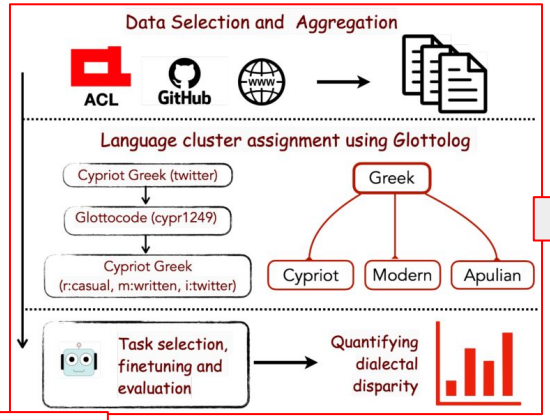
Domain-Task	en-AU	en-IN	en-UK
GOOGLE-Sentiment	0.94	0.64	0.86
REDDIT-Sentiment	0.78	0.69	0.78
REDDIT-Sarcasm	0.62	0.56	0.58
μ	0.78	0.63	0.74

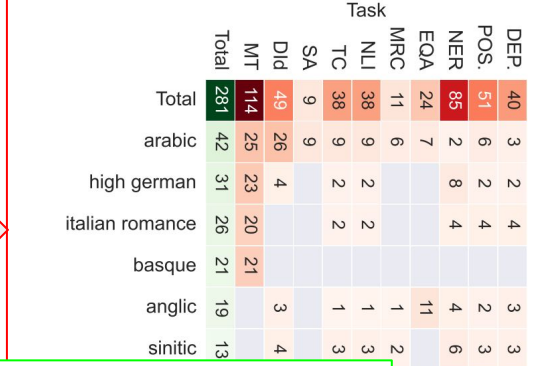
Domain-Task	Mono.	Mult.
GOOGLE-Sentiment	0.86	0.78
REDDIT-Sentiment	0.73	0.76
REDDIT-Sarcasm	0.56	0.60
μ	0.72	0.71

Domain-Task	Enco.	Deco.
GOOGLE-Sentiment	0.86	0.72
REDDIT-Sentiment	0.75	0.75
REDDIT-Sarcasm	0.60	0.55
μ	0.74	0.67

Multi-Dialect Data

Case Study - NLU & NLG - DIALECTBENCH [Faisal et al., 2024]





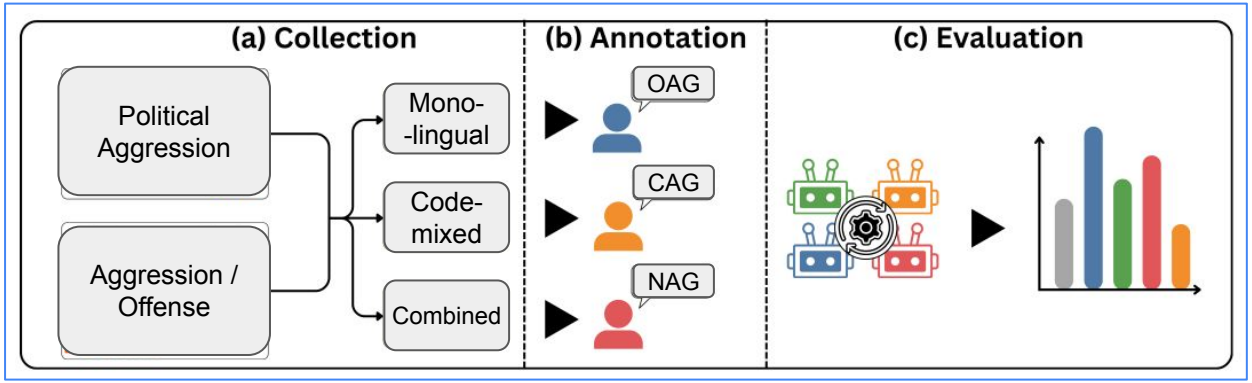
Category	Task	num. cl.	num. var.	avg. score
Structured prediction	DEP parsing	16	40	64.3
	POS tagging	17	51	72.1
	NER	27	85	70.1
Sequence classification	NLI	15	38	64.2
	TC	15	38	77.7
	DId	6	49	67.0
	SA	1	9	80.3
Question Answering	MRC	4	11	40.9
	EQA	5	24	74.2
Generation	MT-dialect	12	73	25.2
	MT-region	2	41	33.0

1. Dependency parsing (DEP parsing)
2. Parts-of-speech tagging (POS tagging)
3. Named entity recognition (NER)
4. Dialect identification (DId)
5. Sentiment analysis (SA)
6. Topic classification (TC)
7. Natural language inference (NLI)
8. Multiple-choice machine reading comprehension (MRC)
9. Extractive question answering (EQA)
10. Machine translation (MT)

Task	Metric
DEP parsing	UAS
POS tagging	F1
NER	F1
DId	F1
SA	F1
TC	F1
NLI	F1
MRC	F1
EQA	Span F1
MT	BLEU

Code-mixed Data

Case Study - NLU -
Aggression Detection



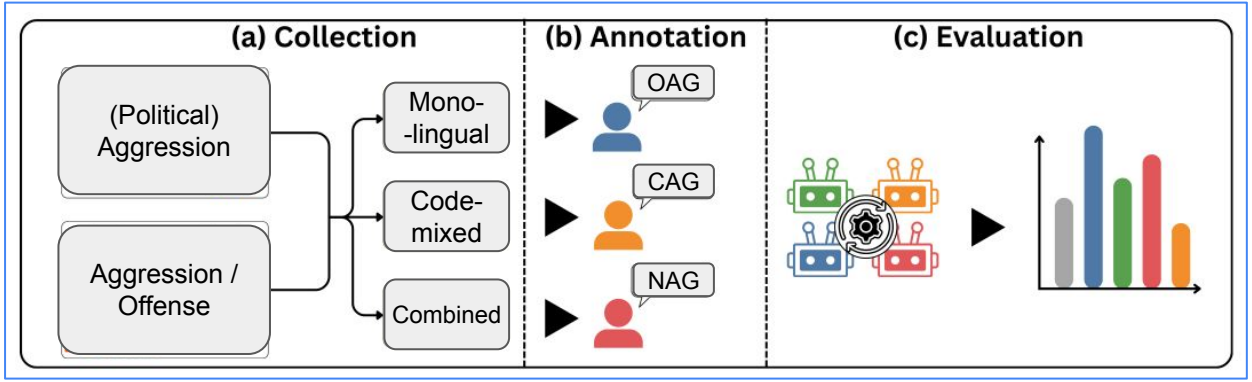
- Two datasets on detection of Aggression on social media
  - Aggression Detection [Political domain] [Rawat et al., 2023]
  - Multi-task Aggression and Offense Detection [Nafis et al., 2023]
- Modelled as 3-class classification problem
- **Challenges**
  - *Crawling* User-generated Content (UGC)
  - *Filtering* code-mixed data
    - Manual filtering *vs.* LID-based thresholding
  - *Pre-training with code-mixed UGC data* helps task performance
  - Nuanced annotation for '*covertly aggressive*'
  - **Cross-dataset performance**: UGC-specific challenge?
    - D1: facebook comments
    - D2: tweets from X

	TRAC (D1)	Ours (D2)	Combined (D3)	Code-Mixed	Non Code-Mixed
$BERT_{base}$	67.17±0.53	58.89±2.42	65.44±0.70	66.40±1.13	63.84±1.42
$RoBERTa_{base}$	69.05±0.57	**66.66±3.82**	66.85±1.23	65.16±2.06	65.11±1.08
$ALBERT_{base-v2}$	66.03±0.89	54.61±4.28	64.71±0.76	62.15±3.89	59.97±2.60
$XLM\text{-}RoBERTa_{base}$	67.73±2.02	61.08±2.21	62.88±3.08	64.52±2.56	60.97±2.73
$MURIL_{base}$	66.64±1.08	60.62±2.00	65.47±0.83	66.71±1.37	62.33±0.88
$XLM\text{-}RoBERTa_{large}$	68.00±1.29	66.38±1.84	**67.95±1.37**	67.83±2.52	64.92±0.97
Hing-BERT	**69.37±0.96**	62.41±3.02	67.48±1.91	**68.50±1.35**	65.13±1.62
Hing-mBERT	67.41±1.06	57.65±2.36	65.70±0.66	65.84±1.71	**65.84±1.40**
HingRoBERTa	68.85±1.28	64.81±2.79	66.95±1.43	68.36±1.71	63.11±1.85

Models	D1 –>D2	D2 –>D1
$BERT_{base}$	48.82±2.55	50.55±1.33
$RoBERTa_{base}$	46.29±3.60	55.33±1.53
$ALBERT_{base-v2}$	46.32±2.58	47.14±1.23
$XLM\text{-}RoBERTa_{base}$	47.32±2.28	52.53±1.19
$MURIL_{base}$	48.77±3.42	52.49±0.68
$XLM\text{-}RoBERTa_{large}$	47.67±2.84	55.77±0.98
HingBERT	47.08±2.38	54.34±1.12
Hing-mBERT	43.06±3.38	52.09±1.87
Hing-RoBERTa	49.30±3.43	52.12±0.71

Code-mixed Data

Case Study - NLU -
Aggression Detection



- Two datasets on detection of Aggression on social media
  - Aggression Detection [Political domain] [Rawat et al., 2023]
  - Multi-task Aggression and Offense Detection [Nafis et al., 2023]
- 3-class for Aggression, *binary classification for offense*
- *Cross-dataset performance in case of binary classes!*
- Erroneous predictions for sarcastic and code-mixed data

	Aggression Detection			Offensive Language Detection		
PTLM	**Monolingual**	**Code mixed**	**Combined**	**Monolingual**	**Code mixed**	**Combined**
BERT$_{base}$	63.58±0.51	65.22±0.77	64.98±0.28	60.99±0.43	61.94±0.14	62.05±0.25
RoBERTa$_{base}$	**66.63±0.12**	65.42±0.61	62.13±0.89	**63.46±0.75**	62.06±0.48	60.21±0.30
XLM-R$_{base}$	65.49±0.73	66.85±0.22	**67.87±0.05**	61.24±0.31	64.42±0.02	65.41±0.73
HingRoBERTa	64.01±0.53	**66.94±0.53**	66.47±0.53	61.92±0.26	**64.97±0.13**	**65.45±0.21**
Bernice	63.49±0.15	61.13±0.43	62.75±0.82	60.88±0.57	59.01±0.38	60.58±0.16

Tweet	**GT**	**M1**	**M2**	**M3**	**Error Type**
As per Zee News 405 for seats for BJP in UP. Total constituency is 403. Two seats given by Zee News on free of cost.	**CAG**	NAG	NAG	NAG	**Sarcasm**
Finally paused the video . It's so nice now lol	**CAG**	NAG	NAG	NAG	**Sarcasm**
Do you know Malda. ??	**CAG**	NAG	CAG	NAG	**Short sequence**
Oh really	**CAG**	NAG	NAG	NAG	**Short sequence**
ek problem hai Main parents ke saath nahi dekh payunga.	**NAG**	CAG	NAG	NAG	**Code-Mixing**
Jay hind Pakistan me jabrdast Hamla Kare Hmari Sena jbab dena jaruri h	**CAG**	NAG	CAG	CAG	**Code-Mixing**

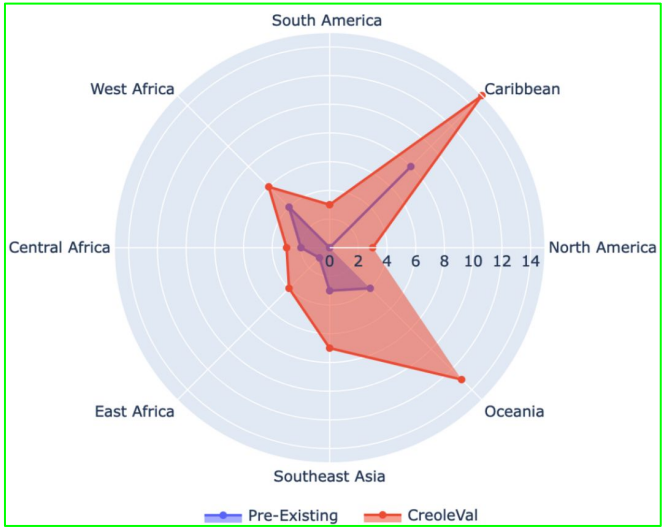
	Aggression Detection		Offensive Language	
	D1–>D2	**D2–>D1**	**D1–>D2**	**D2–>D1**
BERT$_{base}$	55.63±0.21	52.98±0.56	48.69±0.11	46.49±0.53
RoBERTa$_{base}$	52.13±0.74	50.99±0.47	46.02±0.31	43.64±0.49
XLM-R$_{base}$	**56.81±0.84**	55.33±0.60	50.94±0.55	49.27±0.75
HingRoBERTa	56.29±0.71	54.04±0.10	**51.51±0.28**	49.01±0.24
Bernice	52.05±0.87	49.65±0.57	46.16±0.18	45.88±0.05

Creole Data

Case Study - NLU & NLG -

CreoleVal [Lent et al., 2024]

- CreoleVal: Multilingual Multitask Benchmarks for Creoles.
- 28 Creole languages covering 8 tasks spanning NLU and NLG.
- Possible through interdisciplinary collaboration with academics, industry, & community-leaders



- **Single repo unifying pre-existing tasks for Creoles and new ones!**
- NLU results demonstrates the *need for improved transfer methods* (Relation Classification results below)

Dataset	Sent. Enc.	bert-base-multilingual-cased				xlm-roberta-base			
	Rel. Enc.	Bb-nli	Bl-nli	Xr-100	Xr-b	Bb-nli	Bl-nli	Xr-100	Xr-b
Dev(en)		59.63±3.48	76.15±1.59	63.47±1.75	62.15 ±1.65	46.76±2.58	50.58±2.08	49.11±2.51	49.04±1.49
bi		28.01±2.42	25.61±3.92	27.66±5.45	25.96±3.80	18.81±4.04	9.62±0.78	19.42±4.51	14.79±1.77
cbk-zam		20.06±5.88	20.85±6.03	17.67±6.68	17.39±6.45	27.08±6.86	18.48±6.83	18.50±2.77	20.32±2.73
jam		26.97±5.87	15.65±5.00	20.07±5.93	23.98±7.24	10.62±1.27	9.42±5.71	9.06 ±1.70	10.22±0.92
tpi		23.57±4.17	22.90±2.97	22.86±8.13	21.42±5.96	9.36±3.77	11.64±5.54	8.31±8.07	8.48±4.78
AVG		**24.65**	21.25	22.06	22.19	**16.47**	12.29	13.82	13.45

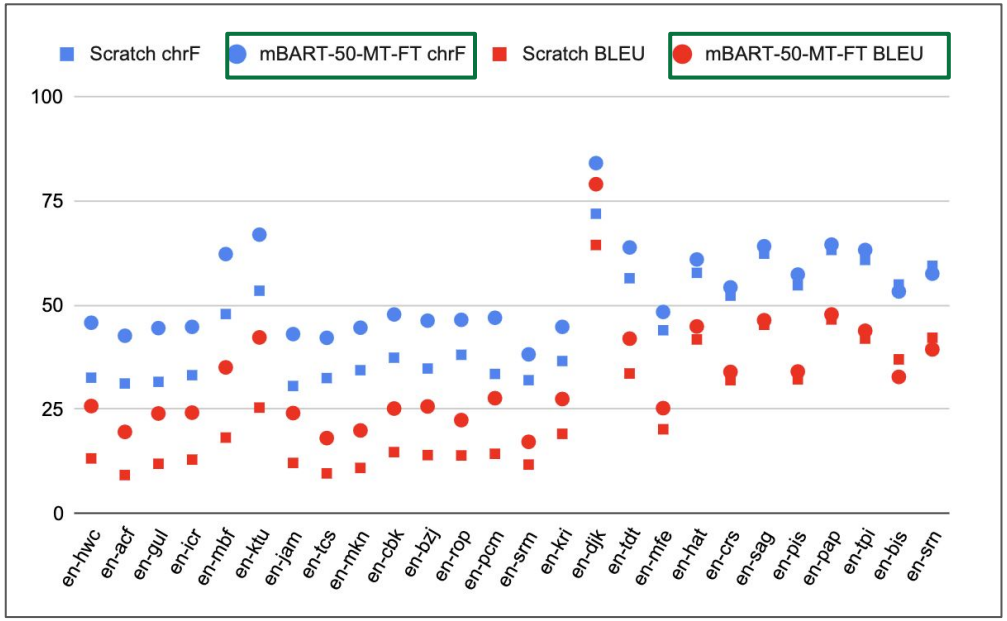
Creole Data

Case Study - NLU & NLG -

CreoleVal [Lent et al., 2024]

- **Religious domain**
  - chrF scores between 30 – 80
  - BLEU scores between 10 – 75
  - Fine-tuned model > Scratch
- New **MIT-Haiti Corpus (educational domain)**
  - The fine-tuned Creole M2M performs surprisingly well cross-domain
  - Versus OPUS-MT model, which are considerably lower than on previous benchmarks
    - **42.2 BLEU** and 59.2 chrF on Tatoeba
    - **14.7 BLEU** and 35.8 chrF on MIT-Haiti

Previous benchmarks have been overly optimistic on performance for Creoles!!!



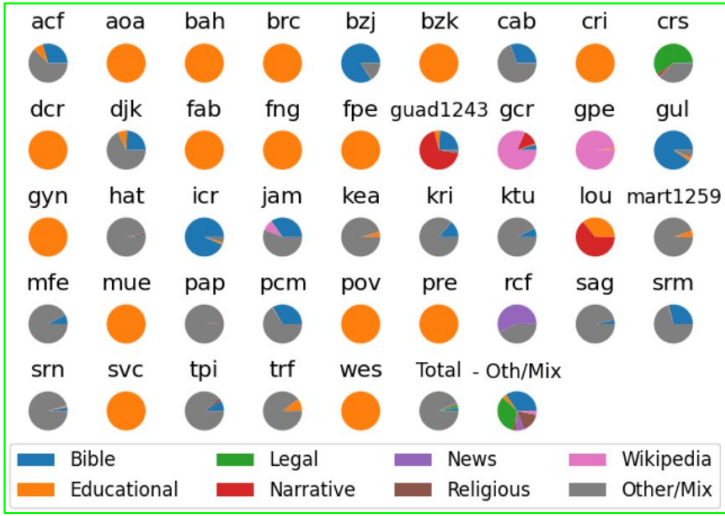
model	source	target	# lines	BLEU	chrF
OPUS	es	ht	102	12.1	32.9
	fr	ht	1,503	11.8	33.5
	en	ht	1,559	14.7	35.8
CreoleM2M	en	ht	1,559	22.0	43.9
	ht	en		18.6	38.1

Creole Data

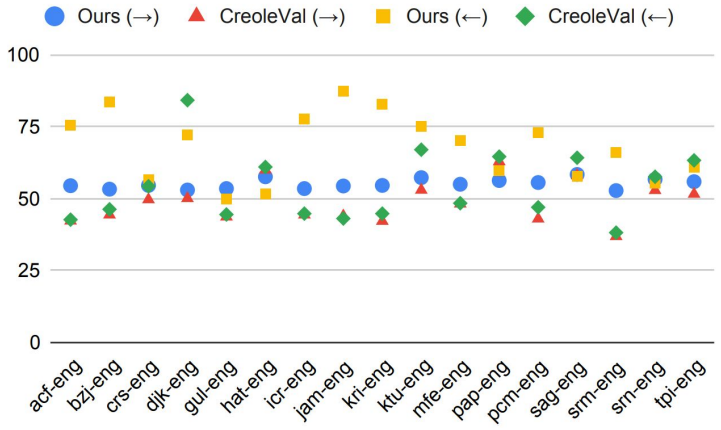
Case Study - NLG -

Kreyòl-MT [Robinson et al., 2024]

MT datasets compiled for 41 Creole languages in 172 translation directions; 11.6M aligned sentences and 3.4M monolingual sentences



Publicly available models w/ SoTA on 23 languages



	Max. prev.	Ours (pub. / all)		Max. prev.	Ours (pub. / all)		Max. prev.	Ours (pub. / all)
acf	15989	4406 / 23916	**gcf**	96	6467 / 6467	**mue**	-	147 / 147
aoa	-	198 / 198	**gcr**	-	1433 / 1433	**pap**	4898029	4968965 / 5363394
bah	-	327 / 327	**gpe**	-	223 / 223	**pcm**	31128	8084 / 47455
brc	-	222 / 222	**gul**	7990	266 / 8831	**pov**	-	480 / 480
bzj	23406	229 / 31002	**gyn**	-	258 / 258	**pre**	-	243 / 243
bzk	-	391 / 391	**hat**	4256455	5715227 / 6023034	**rcf**	-	285 / 285
cab	20879	- / 20879	**icr**	15702	317 / 16774	**sag**	262334	260560 / 535310
cri	-	306 / 306	**jam**	25206	434 / 28713	**srm**	42303	440 / 59053
crs	222613	3186 / 225875	**kea**	129449	132931 / 132931	**srn**	583830	6620 / 615010
dcr	-	189 / 189	**kri**	50438	185 / 66736	**svc**	-	321 / 321
djk	45361	15266 / 68833	**ktu**	7886	175 / 10737	**tpi**	424626	451758 / 925648
fab	-	204 / 204	**lou**	-	1860 / 1860	**trf**	-	1691 / 1691
fng	-	160 / 160	**mart1259**	-	5153 / 5153	**wes**	-	223 / 223
fpe	-	259 / 259	**mfe**	191909	25633 / 233320			

Other Resources

LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation [Paper] [Data]

SentMix-3L, EmoMix-3L, and OffMix-3L: Bangla-English-Hindi Code-Mixed Datasets for Sentiment, Emotion and Offense [Paper 1] [Paper 2] [Paper 3] [Data 1] [Data 2] [Data 3]

GLUECoS: An Evaluation Benchmark for Code-Switched NLP [Paper] [Data]

--

Masakhane Community: Datasets for African Languages [List of datasets]

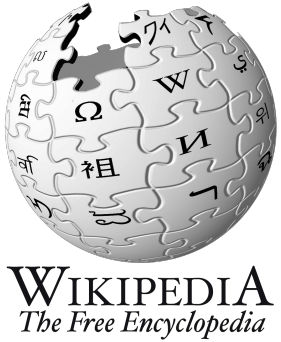
Utilizing Weak Supervision to Create S3D: A Sarcasm Annotated Dataset [Data]

Case Studies

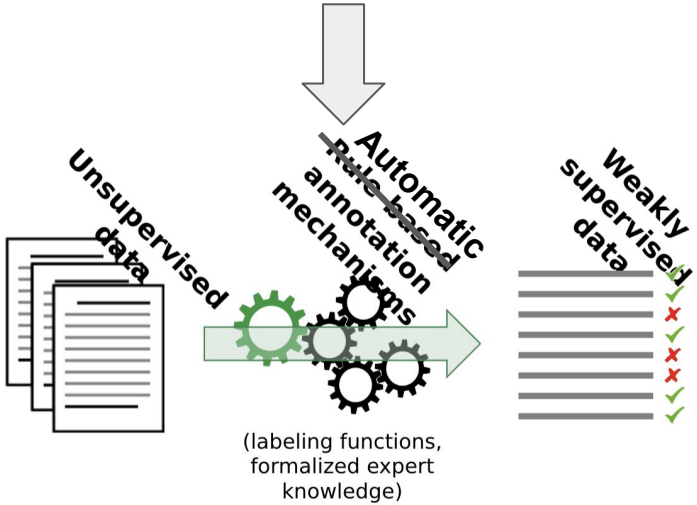
Case Studies: Synthetic Data

- Weak supervision (Lent et al., 2024)
- Rule-based perturbation (Dacon et al 2022 and Ziems et al 2023)

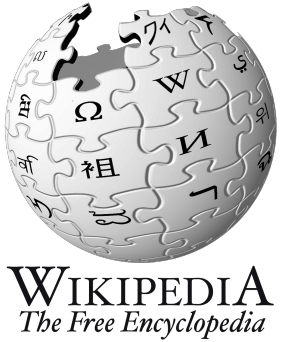
Weak supervision as a pre-processing step



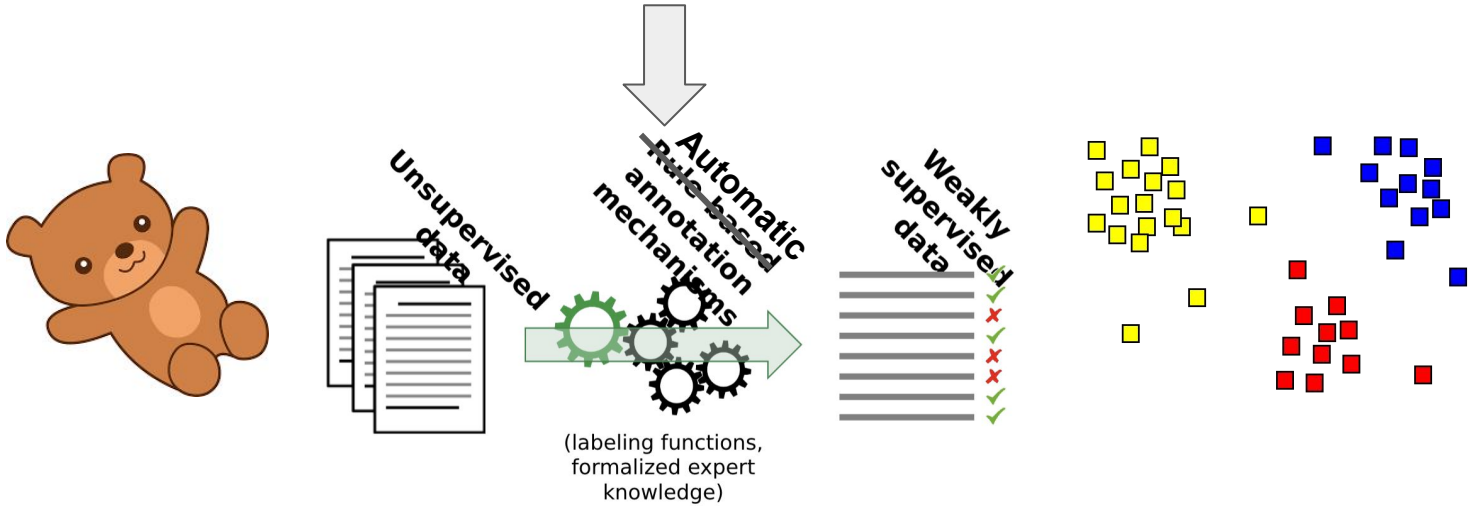
Implicit template: <COUNTRY> a wahn konchri ina <REGION>
Fiiji a wahn konchri ina Uoshania.
Sean Paul a wan Jamaican rappa, singa an dj from Kingston .
Madagiaska a wahn konchri ina Afrika.
Reunioh a wahn terichri ina Afrika.
Sout Sudan a wahn konchri ina Afrika.
Hanova a wah city inna Germany .



Weak supervision as a pre-processing step



Implicit template: <COUNTRY> a wahn konchri ina <REGION>
Fiiji a wahn konchri ina Uoshania.
X
Madagiaska a wahn konchri ina Afrika.
Reunioh a wahn terichri ina Afrika.
Sout Sudan a wahn konchri ina Afrika.
X



Weak supervision as a pre-processing step

Weakly supervised data:
Fiiji a wahn konchri ina [[Uoshania]].
Madagiaska a wahn konchri ina [[Afrika]].
Reunioh [[a]] wahn terichri ina [[Afrika]].
Sout Sudan a wahn konchri ina [[Afrika]].

We can now fix the entity linking very easily by hand

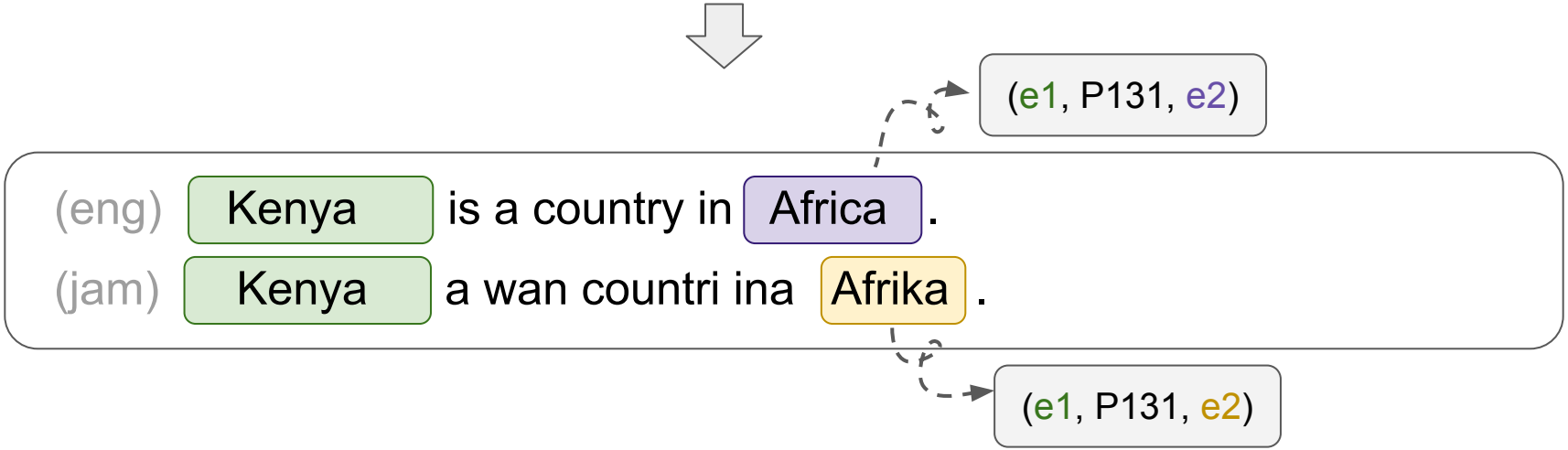
Manual annotation
<COUNTRY> a wahn konchri ina <REGION>
[[Fiiji]] a wahn konchri ina [[Uoshania]].
[[Madagiaska]] a wahn konchri ina [[Afrika]].
[[Reunioh]] a wahn terichri ina [[Afrika]].
[[Sout Sudan]] a wahn konchri ina [[Afrika]].

WIKIDATA

located in the administrative territorial entity (P131)

Quick Speaker Verification → Final Gold Dataset!

Speaker Verification
<COUNTRY> a wahn konchri ina <REGION>
[[Fiiji]] a wahn konchri ina [[Uoshania]].
[[Madagiaska]] a wahn konchri ina [[Afrika]].
[[Reunioh]] a wahn terichri ina [[Afrika]].
[[Sout Sudan]] a wahn konchri ina [[Afrika]].



Facilitates Manual Verification/Correction by Speakers

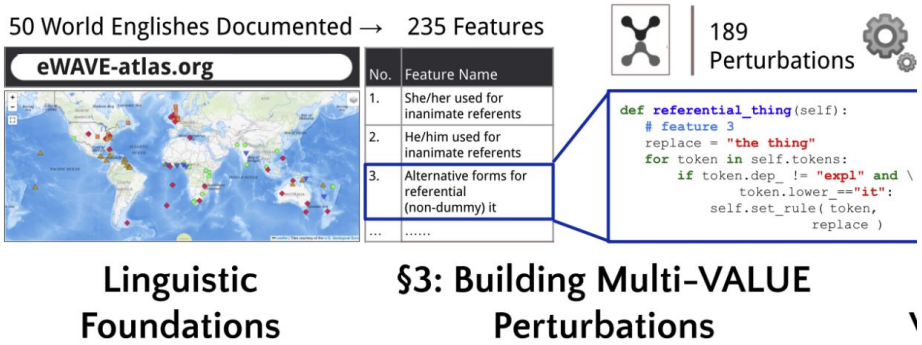
- Wikipedia? Look for templatic sentences.
- Weak supervision to help with manual annotation!
- Grounded in a KB.
- Easy Verification/Correction by Speakers.
- Evaluate zero-shot performance (97 sentences for 4 Creoles).
- With our best models, **F1 scores fall dramatically**.

Dataset	Sent. Enc.	bert-base-multilingual-cased			
	Rel. Enc.	Bb-nli	Bl-nli	Xr-100	Xr-b
Dev(en)		59.63±3.48	76.15±1.59	63.47±1.75	62.15 ±1.65
AVG		**24.65**	21.25	22.06	22.19

Case Study: Rule-based perturbation

Create ‘synthetic’ data using rule-based perturbations on SAE data (Dacon et al 2022)

Multi-VALUE (Ziems et al 2023)



Jamell Dacon, Haochen Liu, and Jiliang Tang. 2022.Evaluating and Mitigating Inherent Linguistic Bias of African American English through Inference. In COLING.
Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023.Multi-VALUE: A Framework for Cross-Dialectal English NLP. In ACL

Case Study 2. Translation

- CreoleVal (Lent et al., 2024)
- **Pros**: Natural, Correct
- **Cons**: Difficulty finding translators, Cost, Cultural Relevance, Translationese

Can’t go wide? Go deep!

- You have some money! Now what?
  - Pick a small dataset
  - Find translators
- Problem: Difficult to find translators

Can’t go wide? Go deep!

- You have some money! Now what?
  - Pick a small dataset
  - Find translators
- Problem: Difficult to find translators
- Translation hiccups
- Solution: Make another localized translation
- Bonus: Cross-cultural NLP!

	Data	#Lang	#Creole	#Lex
mBERT	Wikipedia	104	0	6
XLM-R	CC100	100	0	6
mT5	CC4	101	1	6
CreoleLM (Ours)	custom	128	28	6
mBART-50	custom	50	0	5
Scratch (Ours)	custom	34	26	8

	mBERT	XLMR
English	63.33%	45.00%
Haitian-standard	51.60%	39.16%
Haitian-localized	50.83%	43.33%
Mauritian	49.10%	43.33%

Final thoughts

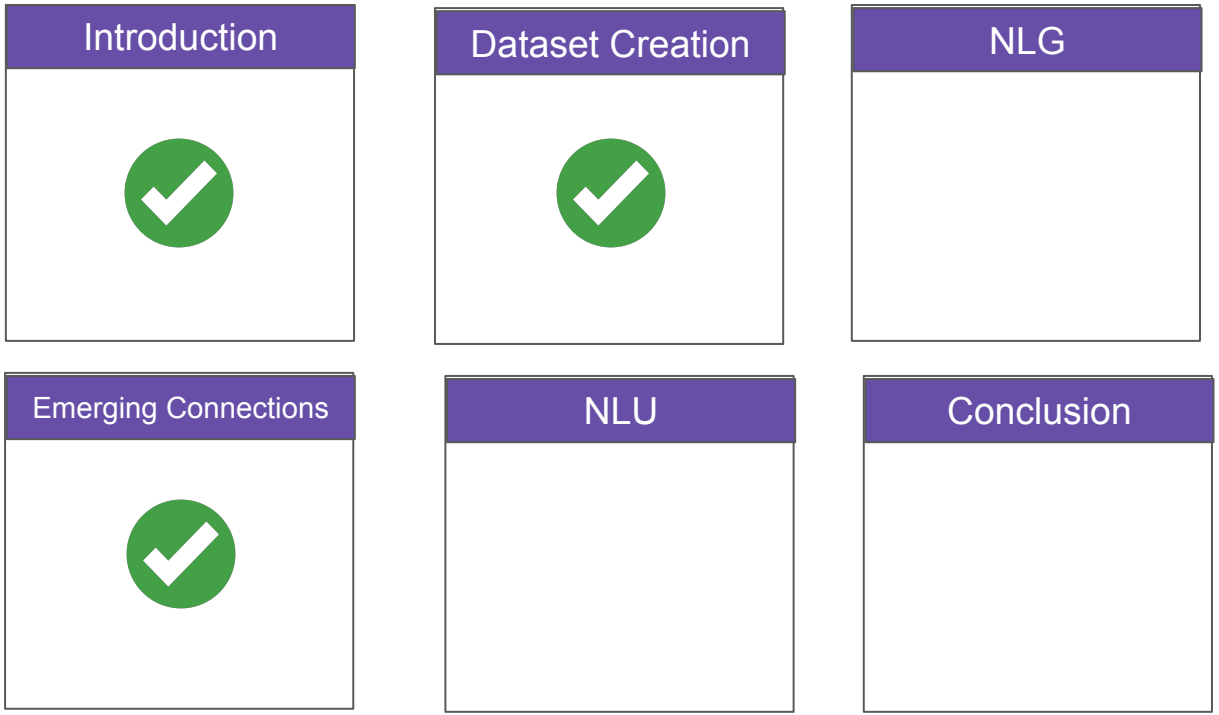
Navigating Pathway to Resource Creation

- Leverage **assisted annotation workflow** for resource creation, especially if you are starting from scratch.
  - Weak Supervision / Zero-shot labelling, Active Learning, Translation Memories,…
- **Manual review** and **validation** at regular intervals ascertains data quality.
- Ensure **domain coverage**, and inspect cross-domain performance.
  - Domain-specificity will only make for a more challenging resource.
- Even **modern encoders** and **decoders** find it challenging to deal with code-mixing, language variety and cross-lingual transfer extremely challenging.
  - Multi-tasking / Multilinguality can help task performance in such cases.
- **Consider releasing your data publicly! :)**

Resource Creation Looks Different from the Inside and Outside

- Many hands make work **lighter**.
  - “Masakhane - Machine Translation for Africa” (∀, Masakhane, 2020)
- “Getting Started” looks different when you are ***not*** a member of the community.
  - “What a Creole Wants, What a Creole Needs” (Lent et al., 2022)
- Ask questions – **Listen & learn**.
  - “Decolonising Speech and Language Technology” (Bird, 2020)
- Act – Don’t Extract! **See how you can give back.**
  - “Ethical Considerations for MT of Indigenous Languages” (Mager et al., 2023)
  - “Must NLP be Extractitive?” (Bird, 2024)

Tutorial Agenda



Q&A