Toloka | TUM | COMMON CRAWL

LREC · 2026 · Tutorial

Low-resource High-Impact

Ekaterina (Katya) Artemova, Laurie Burchell, Daryna Dementieva,
Shu Okabe, Mariya Shmatova, Pedro Ortiz Suarez

Tutorial overview

Two sessions, five case studies

09:00	**Session 1**	
	Part 1: Intro & basics	~30 min
	QA Session 1	
	Part 2: Case studies	
	Case study 1: Pre-training Data Crawling and Filtering	~25 min
	Case study 2: Obtaining Machine Translation System · Part 1	~25 min
10:30–11:00	Coffee break	30 min
11:00	**Session 2**	
	Case study 3: Downstream Tasks System Acquisition	2x~25 min
	Part 3: Expert Interviews	~25 min
	Final QA Session + Closing	

"Data is
the new oil"

— Clive Humby

Data needs in NLP

Raw data

- Used for unsupervised tasks such as LLM pre-training
- Teaches LLMs to generate plausible language
- Usually collected by scraping web sources

Labelled data

- Needed for tasks like sentiment analysis, NER, syntactic parsing
- Targets specific applications (conversational AI, MT) and domains (medicine, law, finance)
- Manually annotated by humans
- Obtainable on platforms such as the self-service Toloka platform

Basics

Data Annotation Basics

How humans turn raw text into training signal — schemas, instructions, quality control.

Example

Labelling with human

Problem formulation

Detect emotions in tweets.

Annotation schema

Six emotions:
love , anger , fear , sadness , surprise , joy .

Quality control

Overlap of 5 annotations · Majority voting
· Control tasks.

Instruction

Please read each text carefully and classify it based on the dominant emotion it expresses: love, anger, fear, sadness, surprise, or joy.If the tweet does not clearly convey one of these emotions, or if it is purely factual or neutral, mark it as neutral.

> Prompt shown to annotators
>
> `i was ready to meet mom in the airport and feel her ever supportive arms around me.`

Saravia et al., CARER: Contextualized affect representations for emotion recognition. EMNLP 2018.

Reality check

Many applications require complex annotation

Beyond sentiment labels: structured annotation of propaganda spans, discourse relations, claim-evidence links, multi-step reasoning traces. Each layer multiplies effort and agreement challenges.

01 Read article → 02 Identify spans → 03 Classify technique → 04 Validate with peers

Overview

Data annotation pipeline

01

Frame the target task

Problem formulation: Does this text pose a risk of a harm?

02

Conceptualize the problem

Annotation schema defines labels, how they should be applied and how complex cases should be treated.

03

Instruct annotators

Annotation guidelines are provided to the annotators to label raw data and to guide annotation decisions

04

Quality control

Choose an **aggregation rule**, control **inter-annotator agreement**, train a baseline model.

Example

Budget estimation

- Let **X** be the cost of annotating a single text
- Target dataset: 1,000 texts + 10% for control tasks
- X depends on hourly rate and average texts per hour
- Time on labelling depends on task complexity and dataset size

Total budget formula

$$(1{,}000 + 100) \times 5 \times X = 5{,}500 \times X$$

10% control tasks · overlap of 5 annotators per item

Saravia et al., EMNLP 2018.

Strategies

Data collection

From the scratch (Native collection)

Collect and annotate original data directly in the target language — gold standard but most costly.

Translation-based transfer

Translate annotated datasets from a high-resource language while preserving labels.

LLM-generated

Use large language models to generate annotated data directly in the target language.

Most real-world pipelines combine all three — translation bootstraps scale, native data calibrates quality, LLM generation fills rare classes.

Translation transfer

Data translation from high-resource language

Source

Sample in English

annotated corpus

→

MT system

Preserve labels

translate only surface text

→

Target

Sample in LR language

projected annotations

Trade-offs

Data translation from high-resource language

Translate data from a high-resource language using machine translation while preserving annotations.

Benefits

- Lower costs for data management and analysis
- Faster data collection turnaround time
- Existing benchmarks can be extended to new languages

Limitations

- MT errors distort culturally grounded expression
- Content may remain culturally anchored in the source
- MT systems may perform poorly in specialized domains

Synthetic generation

LLM-generated data

Prompting an LLM to produce labelled examples directly in the target language is fast — but the gap between high- and low-resource generation quality is severe.

Prompt

```
Create an example in [LR language]
here…
```

Poor ouptut

LLMs often hallucinate morphology, invent non-words, or fall back to the high-resource neighbor — producing text no native speaker would write.

Trade-offs

Synthetic data generation

Generate data that mimics characteristics and features of real-world data.

↗ Benefits

- Lower costs for data management and analysis
- Faster data collection turnaround
- Fewer security issues

⚠ Limitations

- Poor generation in low-resource settings
- Hard to generate accurate and diverse data
- Validation procedures still required
- Quality of trained classifiers may suffer

Fairness

Understanding bias in synthetic data

Synthetic data can perpetuate, amplify, and introduce biases due to several factors:

- Uneven quality in synthetic subsets, such as poorly generated code-switched or vernacular texts, can impact linguistic minorities.
- Neglecting underrepresented groups in data generation can lead to biased outputs that mainly reflect majority perspectives.

Quality control in human labeling

Three phases that separate a usable dataset from a noisy one.

Three phases

Methods of quality control

Before task performance

- Selection of annotators
- Well-designed instruction
- Onboarding and exams
- Choosing the team

Within task

- Technical platform improvements
- Control tasks
- Motivation (performance-based pricing)
- Anti-bot & cheater tricks

After task

- Data acceptance and working with data
- Feedback loops

Phase 01

Before task performance

Before task · 01

Selection of annotators

- Sourcing & selection
  of the supply of annotators

- Choose the right criteria
  (education, languages, country, etc.)

- Select proper specialization for the task
- Control quality of execution level on your tasks
- Pick annotators with best quality on past projects

Before task · 02

Onboarding and exams

Onboarding task

Similar to the production task, but includes guiding comments that walk annotators through understanding the task.

Exam

A test annotators must pass before starting production tasks. Exam tasks evaluate education level and topic understanding.

Phase 02

Within task performance

Within task · 01

ML methods for human labeling

- Smart matching between tasks (domain, complexity, time, format) and annotators(skillset, performance, preferences)
- Smart matching can be a set of simple heuristics; works better at large scale. Can raise quality and speed by 10–30% depending on the task
- ML-based inspirational seeds generation
- ML-based auto-checks and assistants
- Appeals and feedback loops for our experts
- Automatic deduplication and quality metrics

Within task · 02

Anti-fraud rules

Fraud prevention built into the data pipeline from start to finish to guarantee authentic human effort and expertise.

- Response speed control
- Cursor trajectory check
- CAPTCHA checks
- Identity verification
- Link-visit tracking
- Video-playback check

Within task · 03

Control tasks / honeypots

Tasks with a known correct answer shown to performers to evaluate their performance.

- Distribution of answers in control tasks = distribution in the whole set of tasks
- Should still contain rare answer variants with higher frequency
- Refresh your set of control tasks regularly to avoid bots and cheating
- Automatic control-task generation via annotators
- Tasks with answers of high confidence (aggregation from a large number of annotators)

Phase 03

After task performance

After task · 01

Data acceptance

- Data-quality metrics correlate with model-performance gains — confidence in training data
- Audit by annotation or domain experts
- Aggregation tools for general crowd tasks
- ML-based assessment of dataset quality, including LLM-checks

After task · 02

Inter-annotator agreement

People tend to perceive things subjectively — even professionals.

How to improve it

- Clear annotation guidelines
- Few-shot examples
- Training & calibration
- Measure agreement metrics
- Iterative feedback & revisions

Q&A

Questions?

Part 02

Case studies

Case study 1

Dataset creation through community annotation

Laurie Burchell and Pedro Ortiz Suarez · Common Crawl Foundation

Introduction

Overview

- Today: a case study from our upcoming ACL paper
- Key contribution: **CommonLID** dataset for language identification evaluation
- Aim: share lessons learned about **community annotation**

Motivation

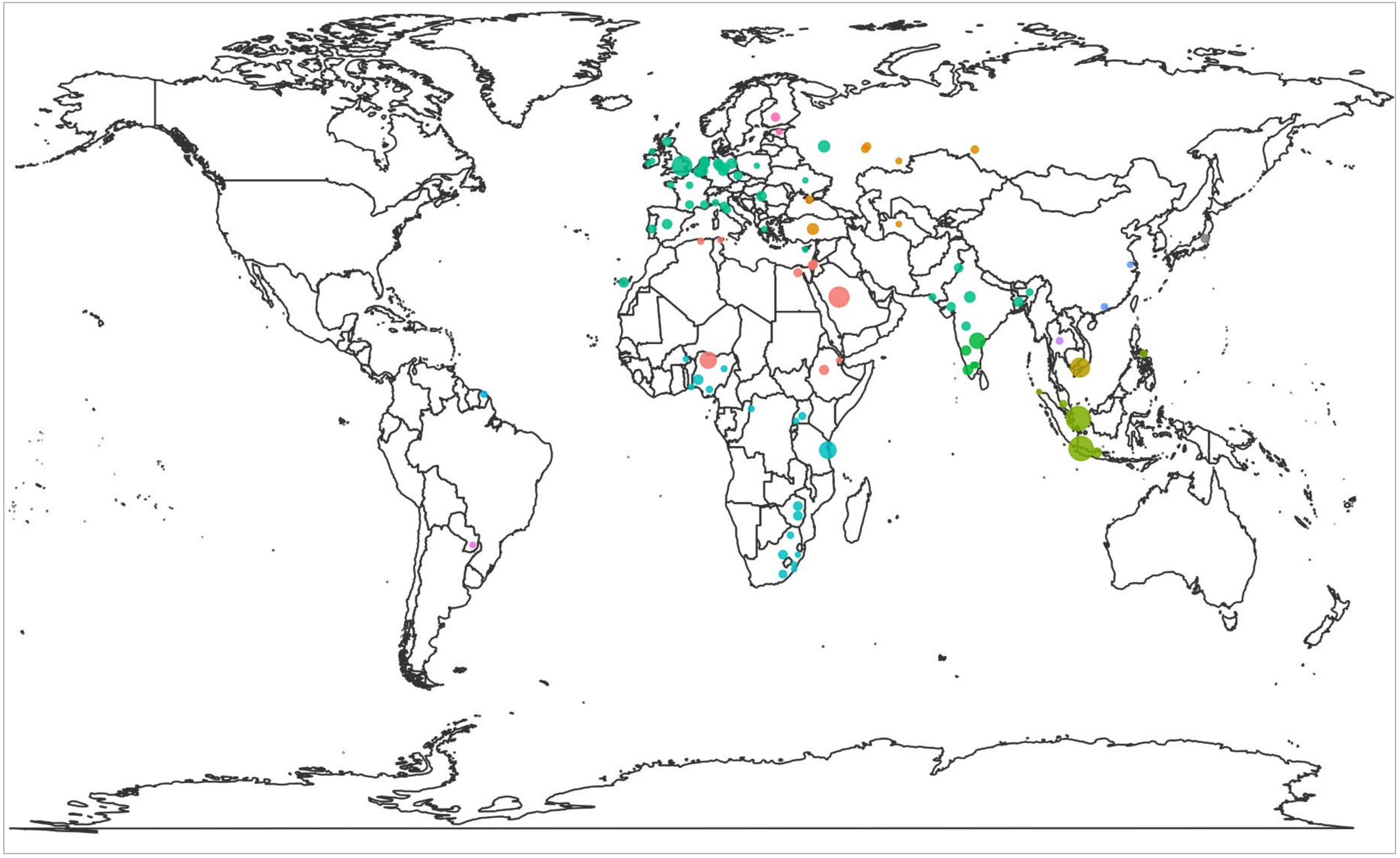
Why CommonLID?

- Language identification (LID) is key for **corpus creation**
- LID at scale is **not solved, especially for long tail**
- Kreutzer et al. (2022): for many languages in web-derived corpora, the majority of sentences were not the right language!
- Test sets are the first step in building better systems, but **few web evaluation datasets** exist

[1] Kreutzer et al., Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. TACL 2022.

CommonLID

A community-driven, human-annotated LID benchmark for the web domain, covering 109 languages.



Statistics

CommonLID summary statistics

373,230

Lines

109

Languages

1–43,189

Lines per class (mean 3,424)

78

Languages with >100 lines

Process

Steps to create CommonLID

1. Sampling from Common Crawl
2. Bespoke annotation platform
3. Participant recruitment
4. Cleaning annotated data
5. Paper and final release

Step one

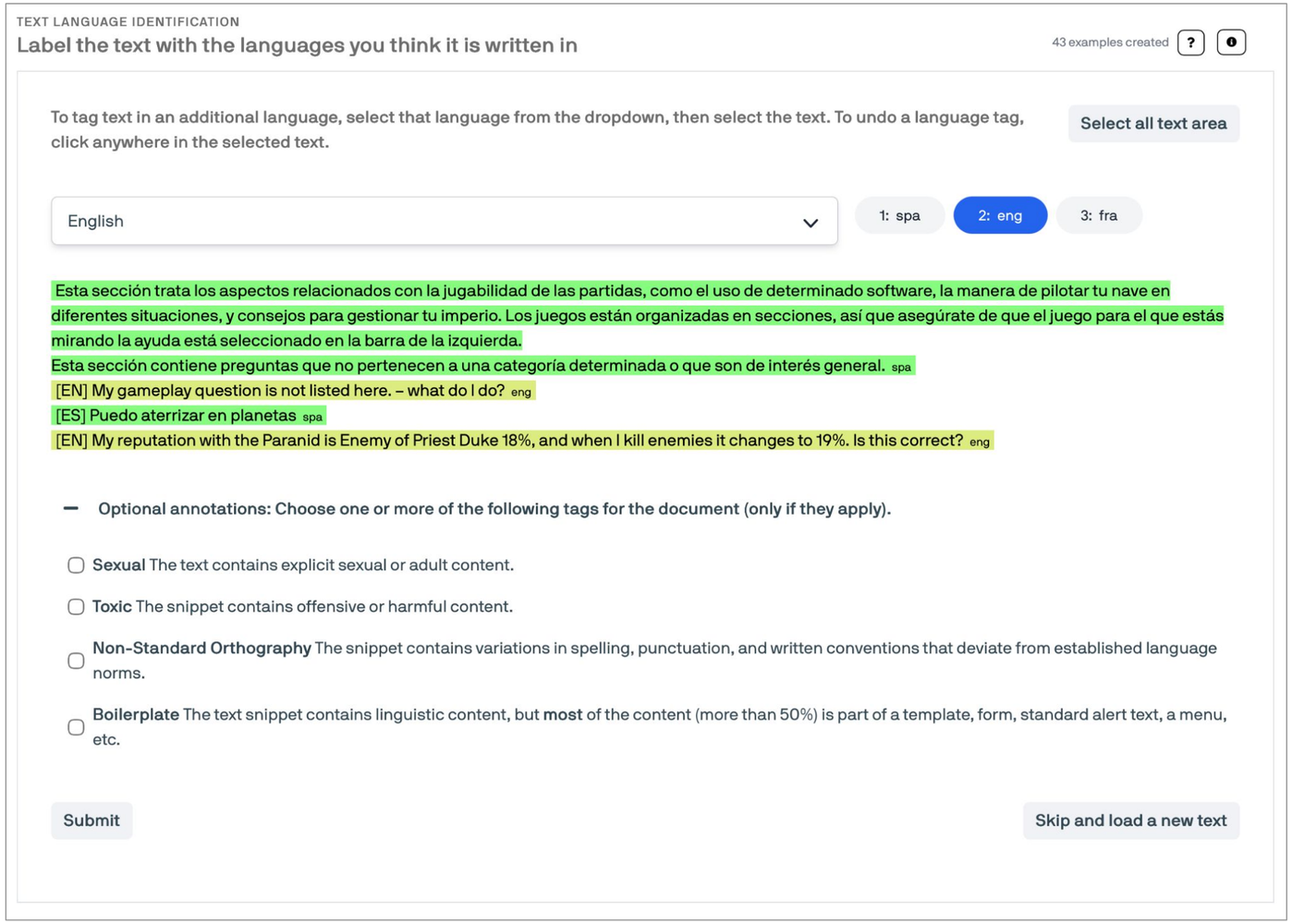
Sampling from Common Crawl

- Three different language identification models to preselect
- Two crawls plus MADLAD (derived from Common Crawl)
- Due to scale, finding relevant content randomly **incredibly unlikely -** need to filter!

Step two

Custom annotation platform

A screenshot of the annotation platform interface. The participant has highlighted the English and Spanish text in the extract in different colours.



Step three

Participant recruitment

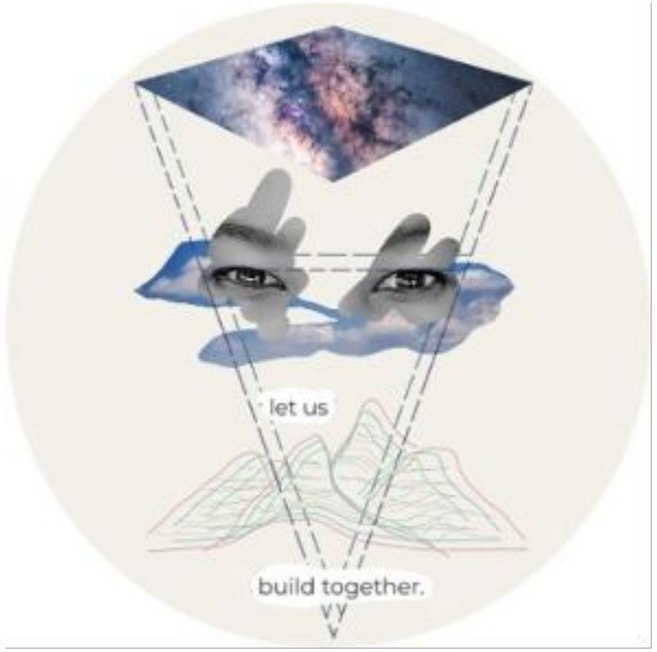
We advertised widely in the NLP community:

- Social media
- NLP Discord servers
- Grassroots NLP organisations

Most participants = researchers wanting to improve NLP for their language(s)
→ engaged, motivated, high-quality annotators

Collaboration

Hackathons with Masakhane and SEACrowd



SEACrowd

Step four

Dataset preparation

- Combination of **automatic** and **manual** steps
- Automatic: deduplication, filtering short lines, English check
- Issue: **label discrepancies**
  - Some labelled at word-level
  - Overuse of “select all”
  - Multiple valid labels (e.g. Arabic dialects)

Processing

Manual audit

- Time consuming but **crucial**
- Check all lines assigned more than one label
  - Often obvious mistake (often English) → remove
  - Sometimes multiple valid labels → keep all
- Audit a sample of all languages
- Not perfect - we ask for community feedback!

Step five

The paper

- **97 authors** was a challenge!
  - Many ARR and ACL forms
  - Lots of admin for all authors
  - Many deadlines with threats of desk rejection 😓
- Co-authors very engaged, so we managed
- Next time: better communication channels and better logistics from the start

Takeaways

Lessons learned

Communication is crucial

- Scalable mailing list
- Annotation instructions
- Regular updates

Appreciate your annotators

- Provide incentives
- Respect their time
- Make it fun

Allow plenty of time for admin

Read more

More information in our paper

Full methodology · Dataset release · Baselines

COMMON CRAWL · ML Commons · Factored · EleutherAI · Johns Hopkins University

Q&A

Questions?

Case study 2

Obtaining Machine Translation System

Obtaining Machine Translation System

Parallel sentence mining for low-resource languages

By Shu Okabe

Motivation

Parallel corpora

A valuable resource for downstream NLP tasks: most notably, **Machine Translation**.

Source · English	Target · German
Please rise, then, for this minute's silence	Ich bitte Sie, sich zu einer Schweigeminute zu erheben
(The House rose and observed a minute's silence)	(Das Parlament erhebt sich zu einer Schweigeminute)
Madam President, on a point of order	Frau Präsidentin, zur Geschäftsordnung
This is all in accordance with the principles that we have always upheld	All dies entspricht den Grundsätzen, die wir stets verteidigt haben

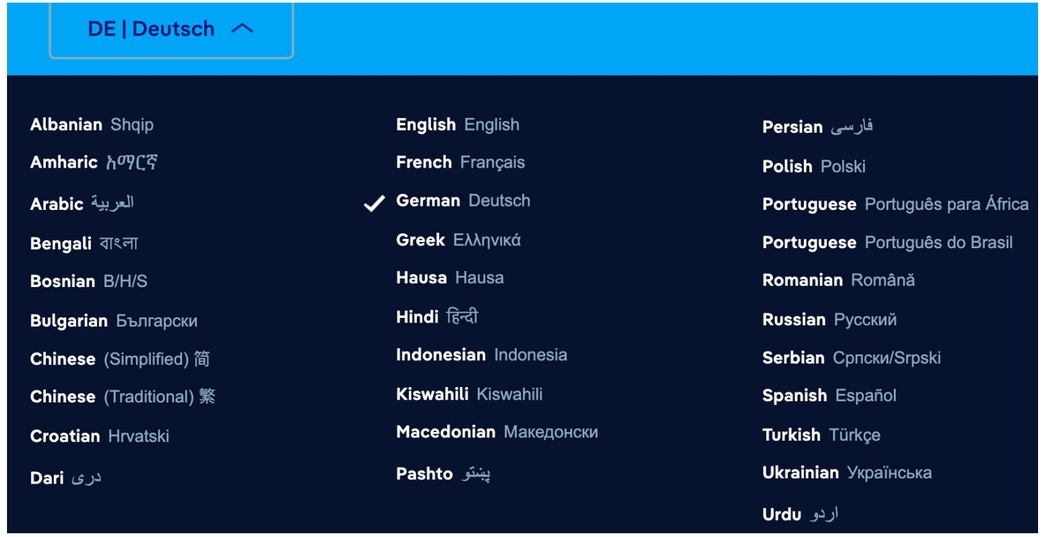
Adapted from the Europarl corpus (European parliament).

Sourcing

Where to find potential parallel sentences?

Multilingual websites

News portals

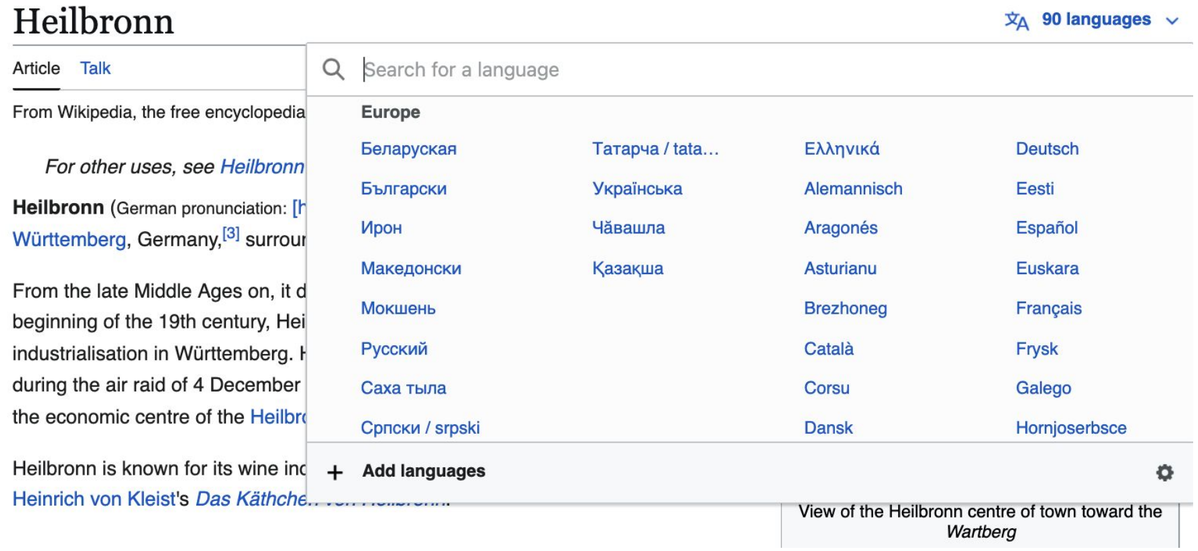


Sourcing

Where to find potential parallel sentences?

Wikipedia

Cross-lingual article alignments



Sourcing · Existing resources

Existing parallel corpora

OPUS

Portal of parallel corpora

CCMatrix / NLLB

Parallel sentence mining on Common Crawl snapshots

Flores

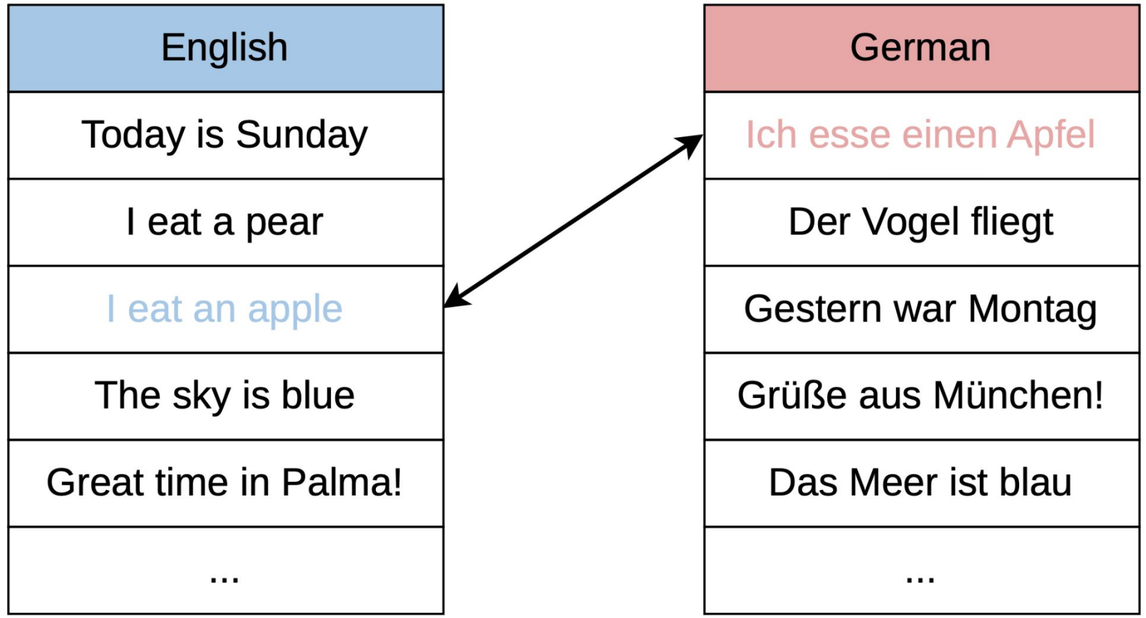
MT evaluation benchmarks covering 200+ languages

Parallel sentence mining

Extracting parallel sentences from
two monolingual corpora.

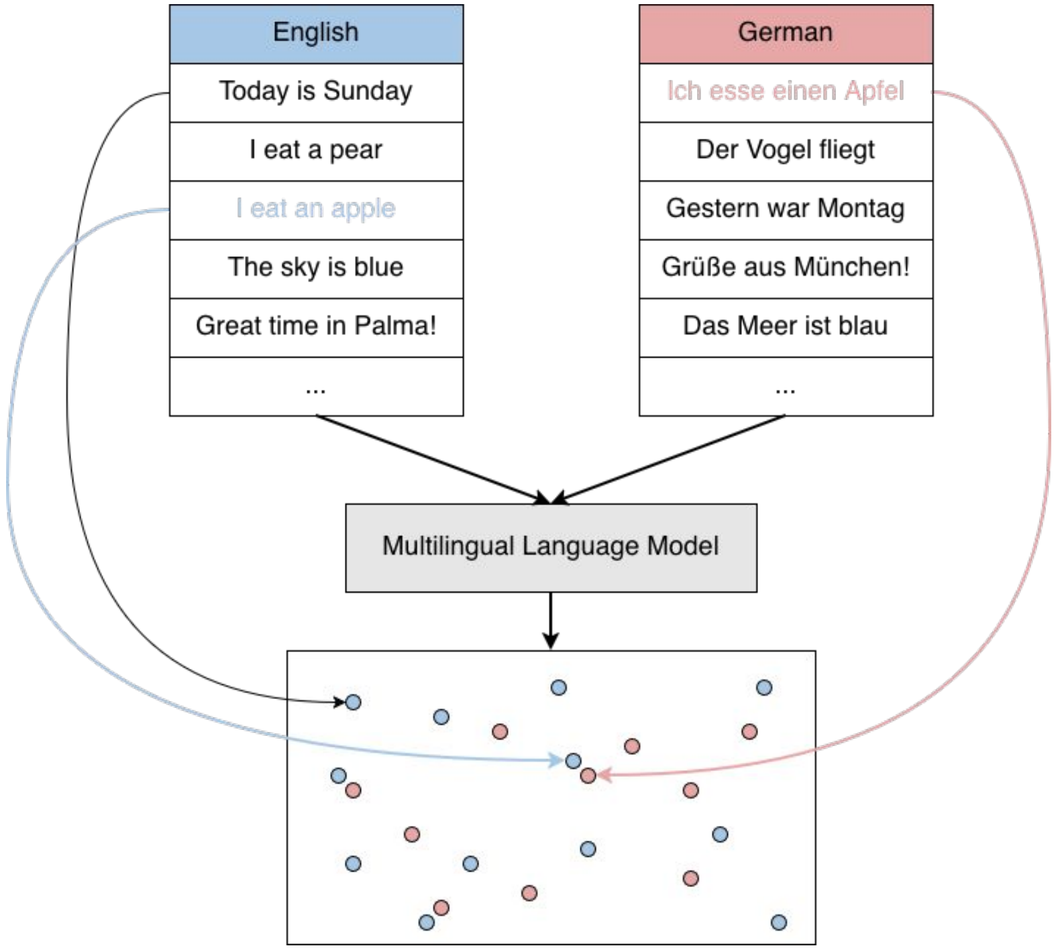
Challenges:

- No guarantee that a sentence has a matching counterpart in the other corpus
- Differences between the two languages



Mining · Step 01

From monolingual sentences to vectors

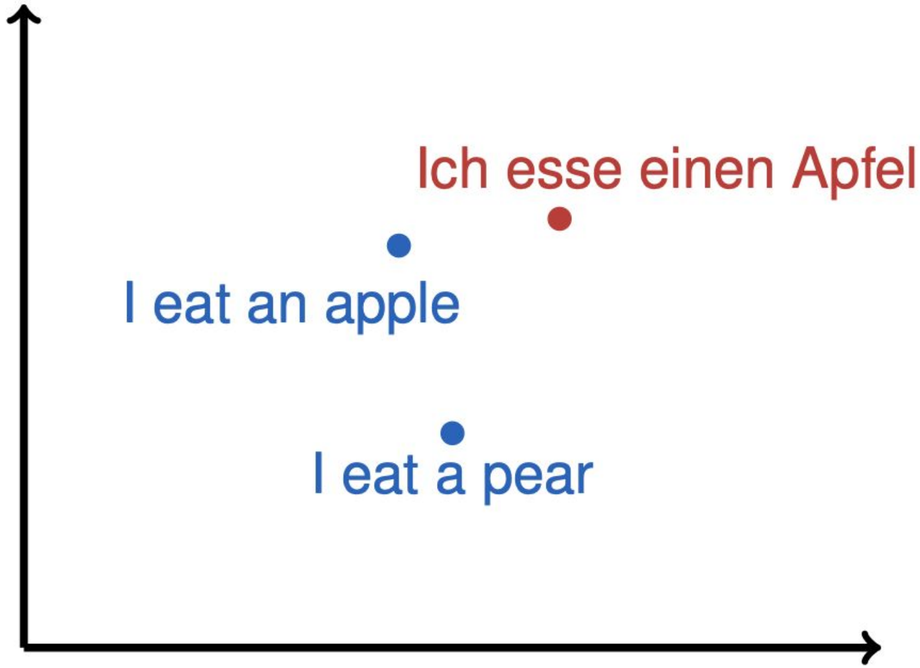


Mining · step 02

Compute similarity between sentences

Similarity between multilingual sentence embeddings as a score

- Cosine similarity
- Improved similarity score based on cosine similarity

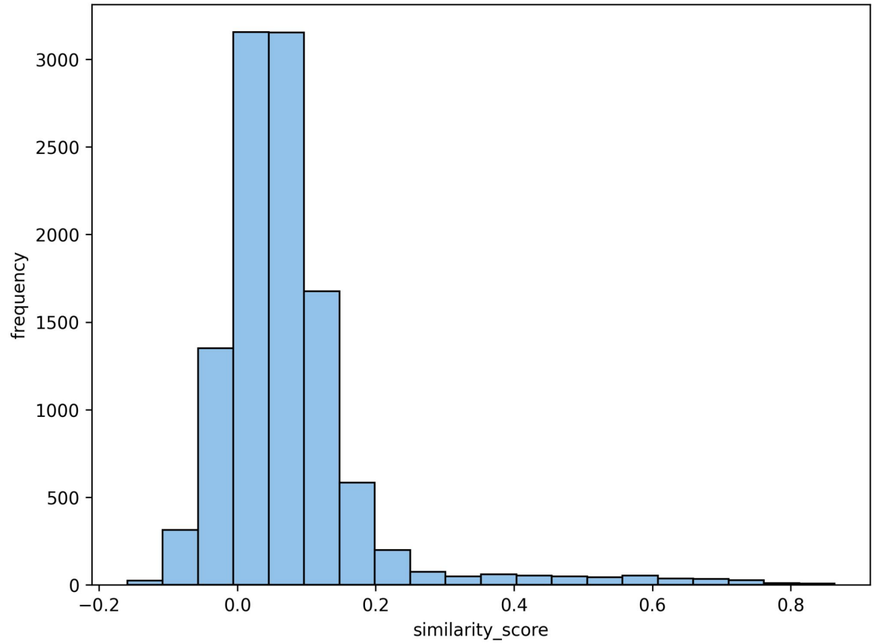


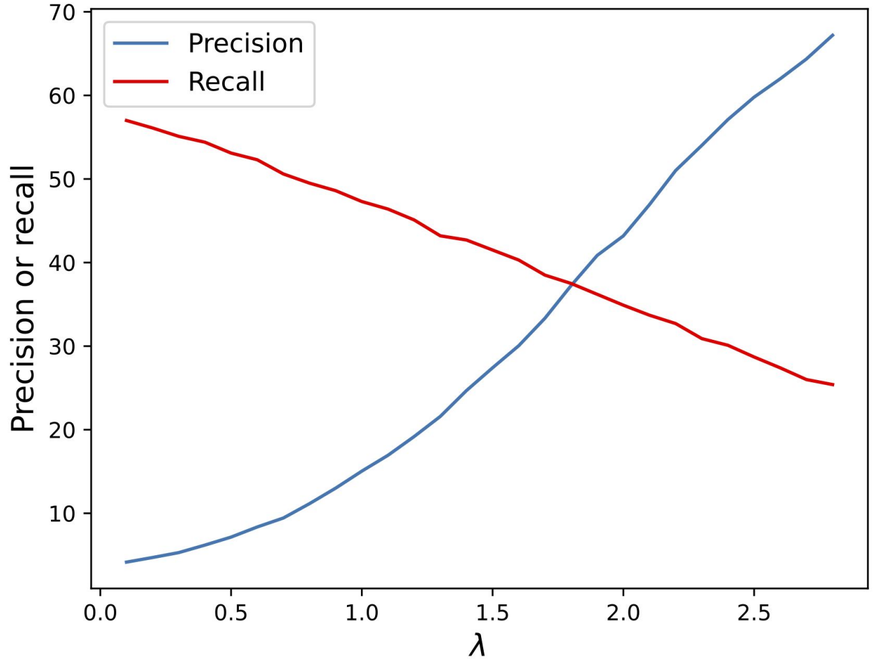

```
sim("I eat an apple", "Ich esse einen Apfel") > sim("I eat a pear", "Ich esse einen Apfel")
```

Mining · Step 03

Filtering sentence pairs

Setting a **similarity threshold**
to separate true parallels from noise.





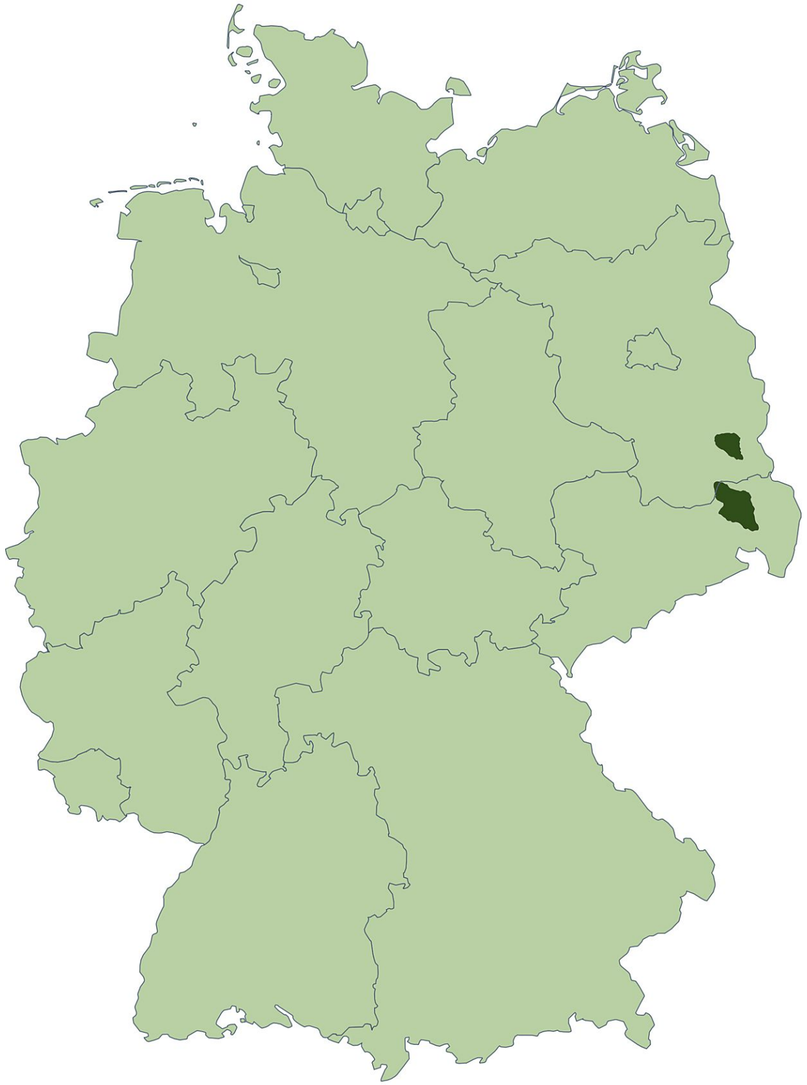
Case study

Low-resource languages: Sorbian languages

Sorbian case study

Upper and Lower Sorbian

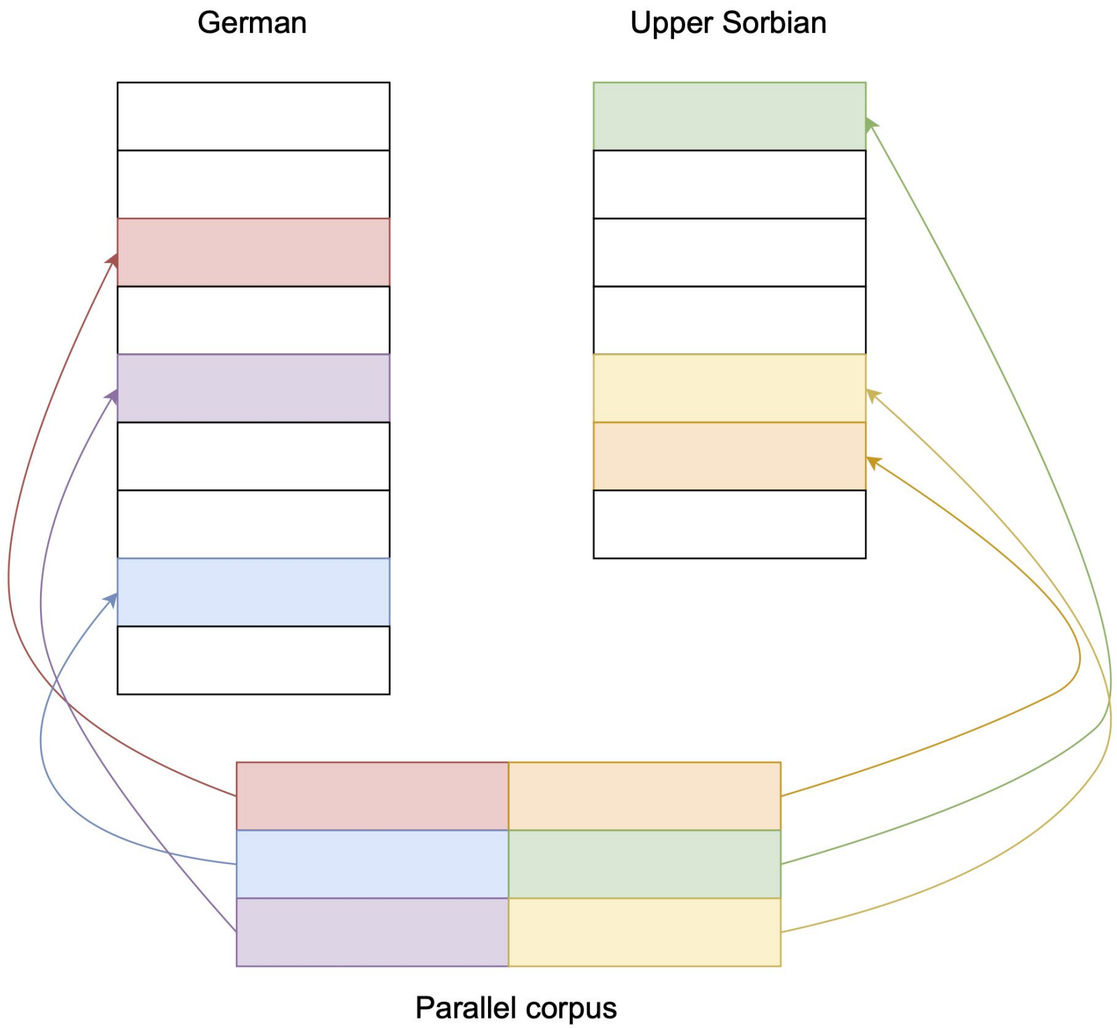
- Two endangered West Slavic languages (ISO codes: hsb and dsb) spoken in Germany
- Previous co-operation with non-profits
  - Shared Task organisation
- Partnership with Witaj Sprachzentrum (Witaj Language Centre) and the Sorbian Institute



Methodology

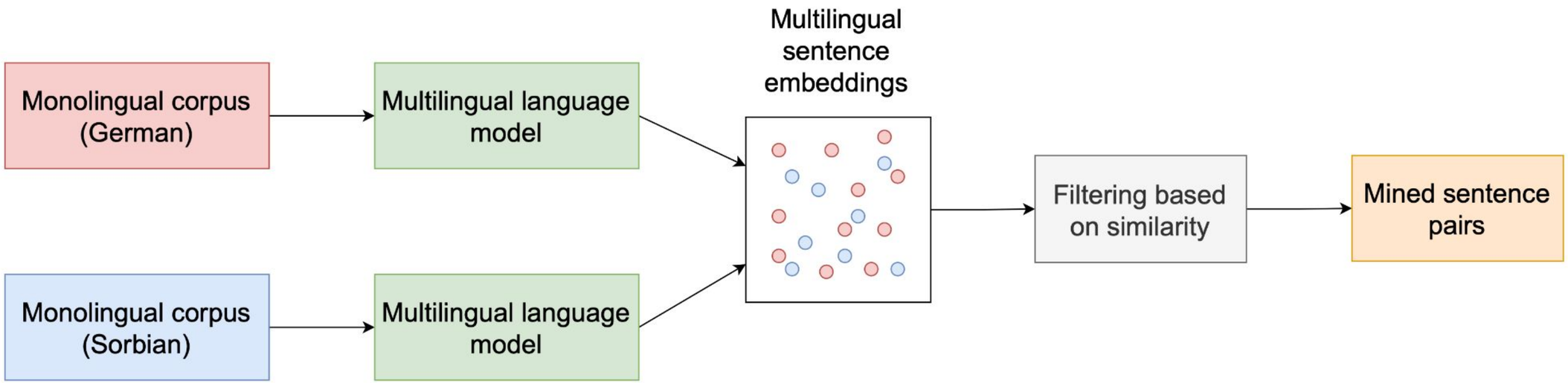
Experimental methodology: corpus creation

- Inject parallel sentences in monolingual corpora
- Build a **BUCC-style corpus** for evaluation



Pipeline

Mining pipeline



Multilingual language models

Multilingual sentence embeddings come in two ways:

- Average of word embeddings in the sentence
- Models trained **directly** to represent the sentence

3 off-the-shelf models

Model	Description	Reference
XLM-R (base)	Baseline multilingual language model	Conneau et al., 2020
Glot500-m	Extension of XLM-R to low-resource languages	Imani et al., 2023
LaBSE	Multilingual sentence-level encoder	Feng et al., 2022

Models

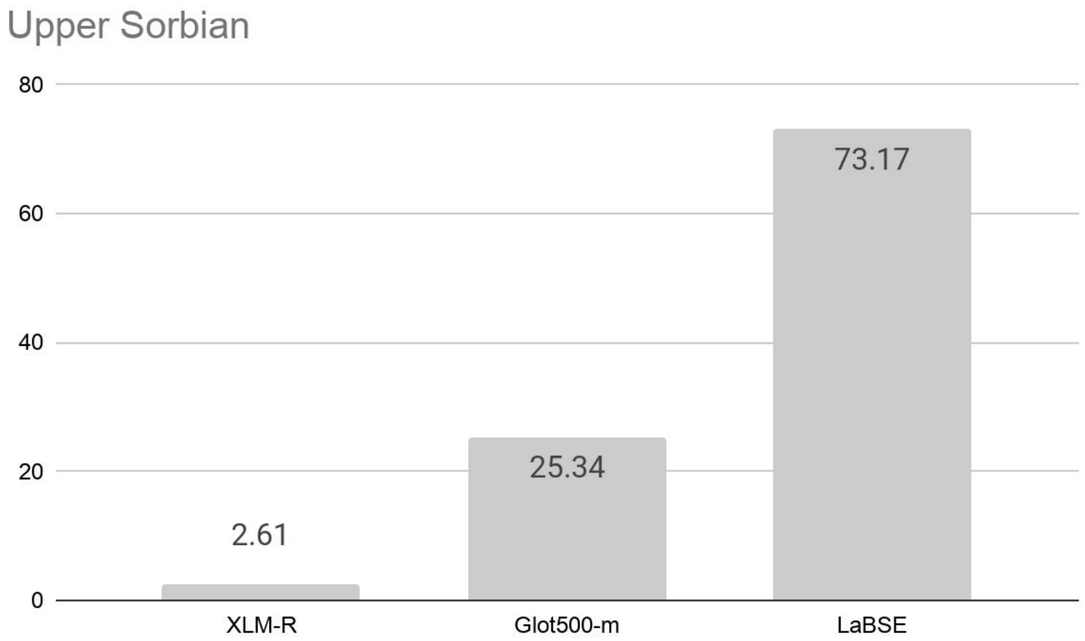
Multilingual language models

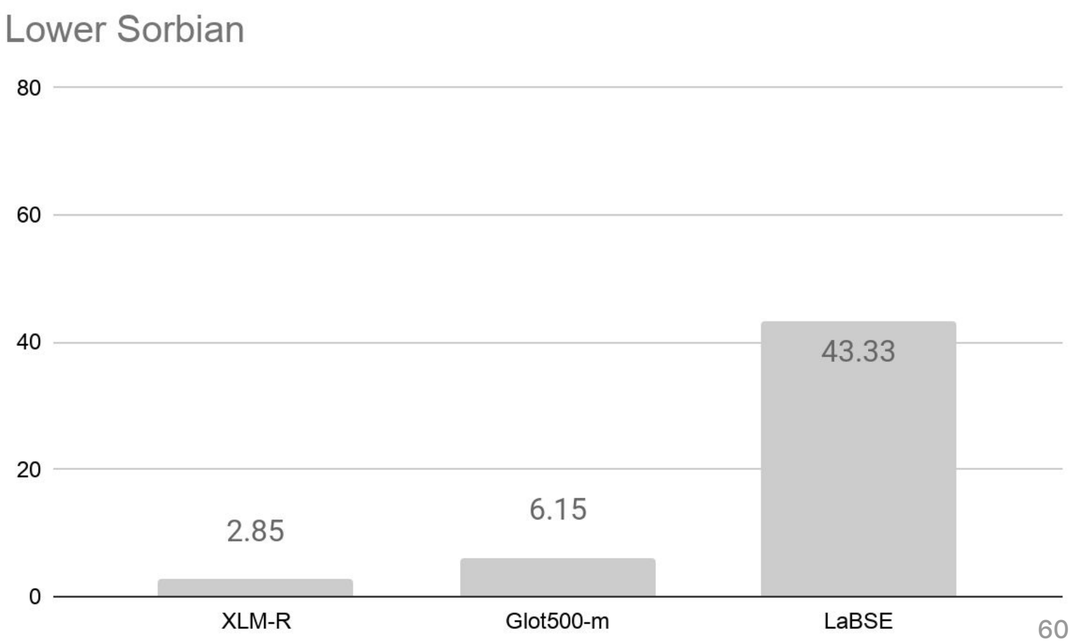
	XLM-R	Glot500-m	LaBSE
German?	✅	✅	✅
Czech & Polish?	✅	✅	✅
Upper Sorbian?	❌	✅	❌
Lower Sorbian?	❌	❌	❌

Results

Mining results for Upper & Lower Sorbian

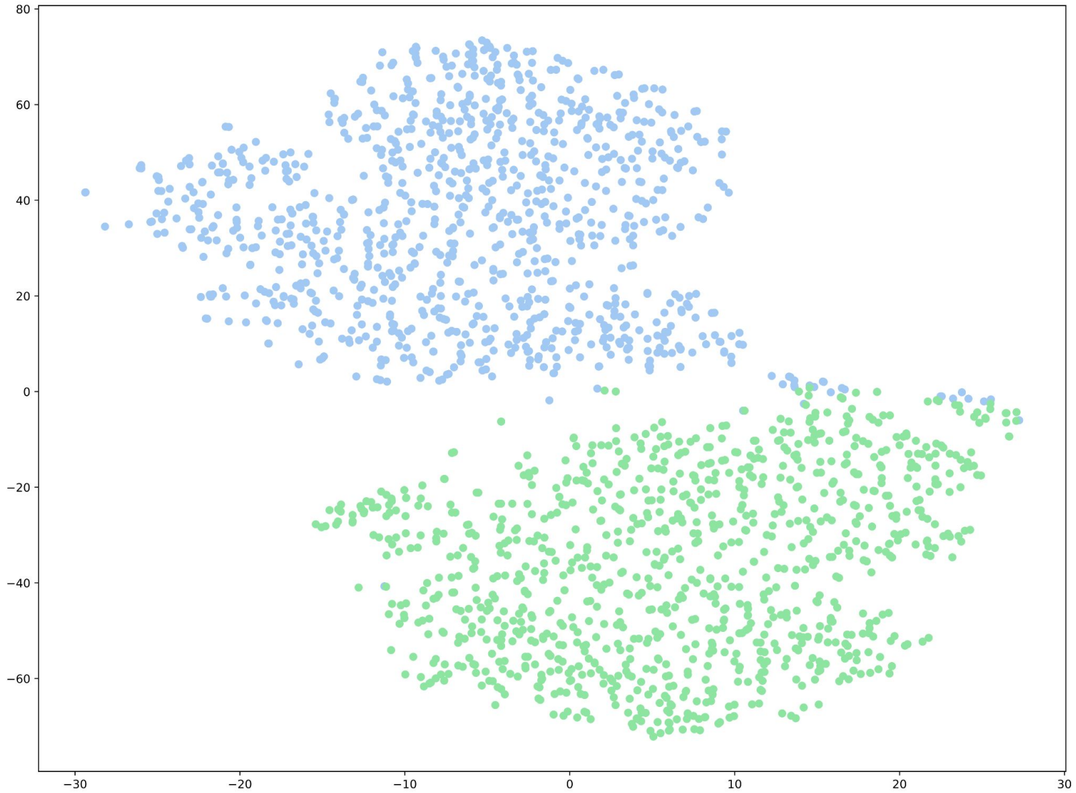
Measuring how well the tool retrieves the true parallel sentence.

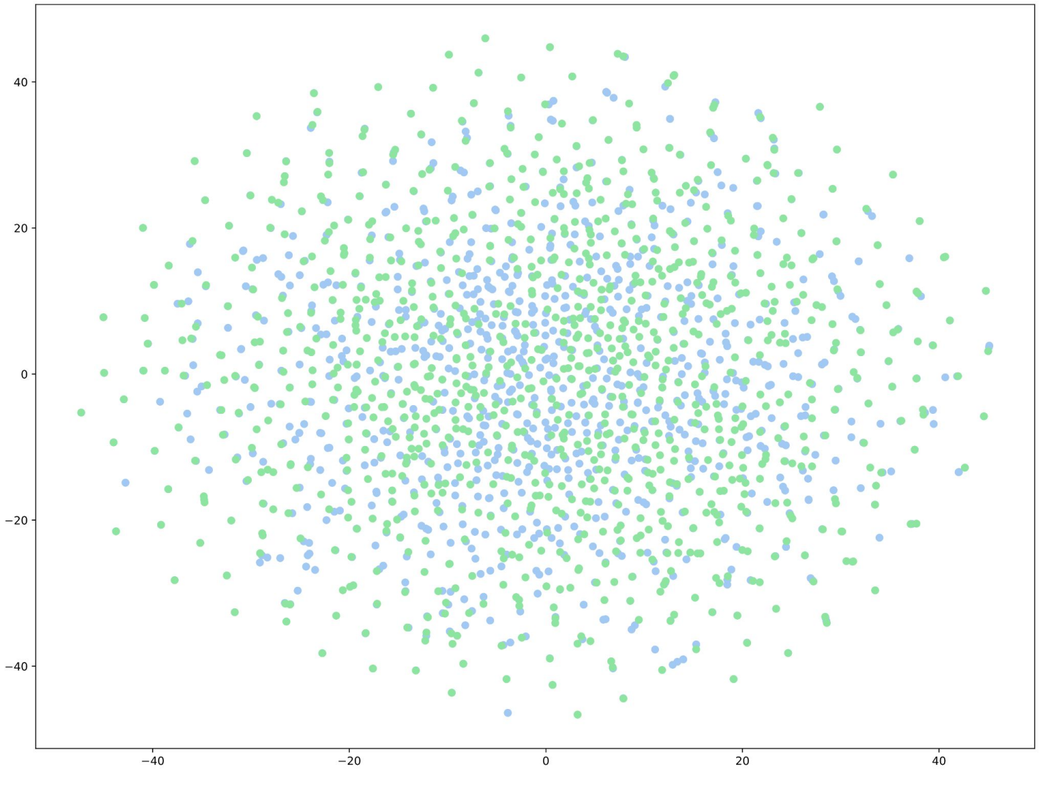




Models

Cross-lingual misalignment

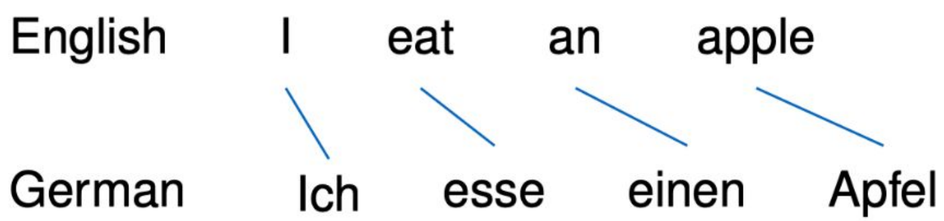




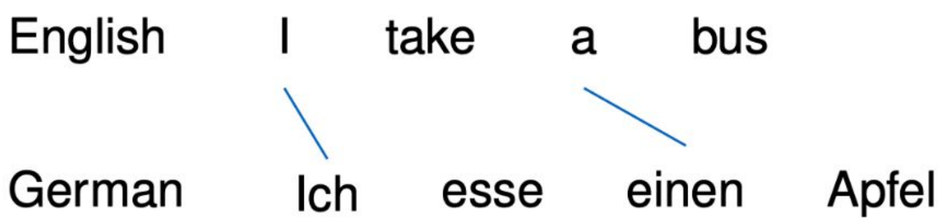
Refinement

Parallel sentence filtering

Alignment post-processing



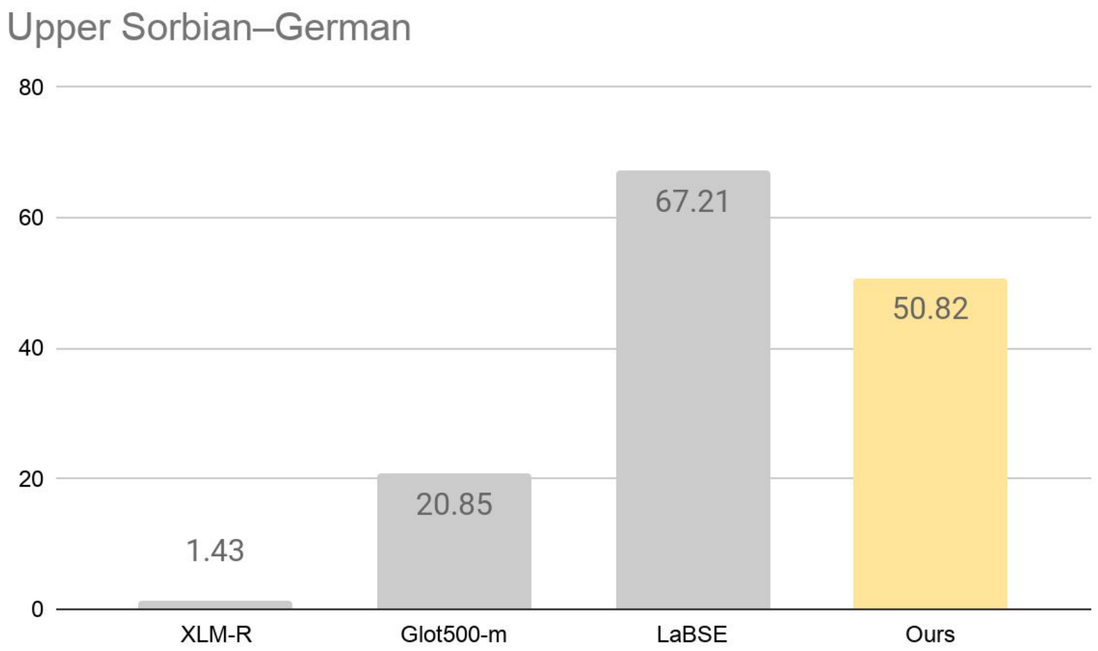
4 alignment links
(alignment score: 100%)

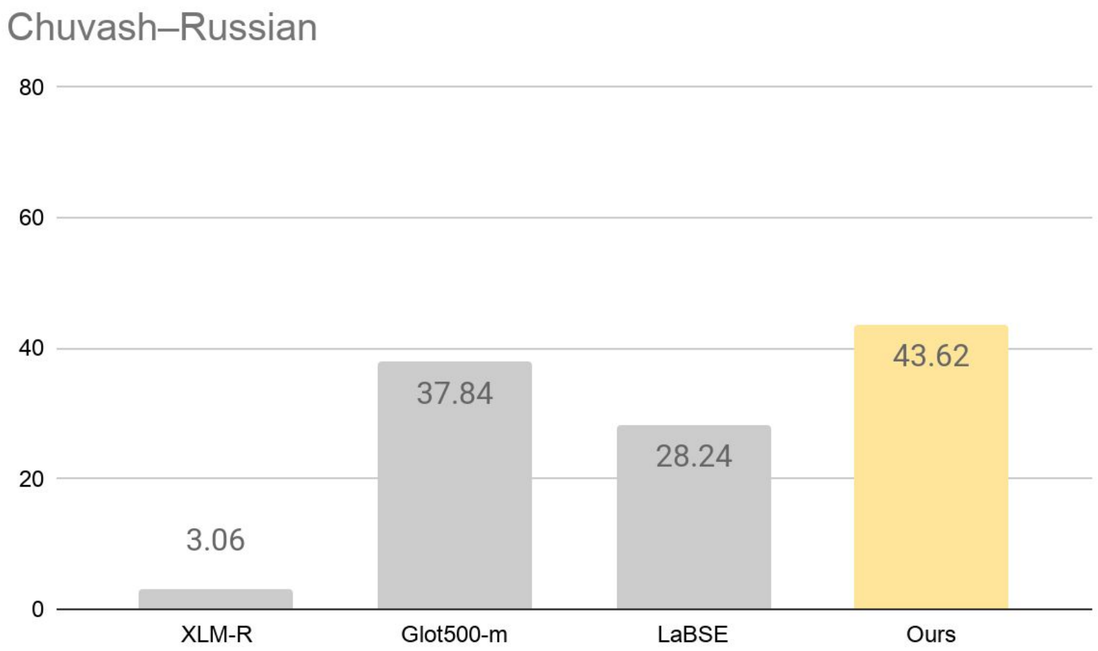


Only 2 alignment links
(alignment score: 50%)

Refinement

Improving cross-lingual alignment





Case study 02

Conclusion

- Parallel sentence mining: Curating **parallel** sentences from **comparable** corpora
- Challenges for mining **low-resource** languages and language pairs:
  - Cross-lingual misalignment
  - Lack of parallel sentences
- Case study on **Upper and Lower Sorbian paired with German**
- Findings:
  - **Monolingual pre-training** is crucial
  - **Cross-lingual transfer** from related languages is effective

Q&A

Questions?

Case Study on Upper and Lower Sorbian

Extension: mining benchmark

WMT2026 Shared Task

Pause

Break

Toloka Platform for data labelling

WMT2026 Shared Task

Use cases For Ukrainian

by Daryna Dementieva

Use case

Text classification when target language is not English, and you have no data

Setup

Let’s assume we have data in English, but want classifiers for Ukrainian

Illustration

Three text classification tasks

Toxicity / Hate speech

Classify text as toxic or non-toxic (or hateful / non-hateful)

Example

```
listen u wikipedia f**s unblock me or i kill u all
```

Formality

Classify text as formal or informal

Example

```
I know i know u seen funnier but it still makes me laff :)
```

NLI · Fluency

Classify text as fluent or non-fluent

Example: contradiction

```
P: Three firefighters come out of subway station.(Premis)

H: Three firefighters playing cards inside a fire station.(Hypothesis)
```

There might be rude texts — these are only examples.

Requires Translation System

Requires emerging abilities of knowledge transfer between languages

Requires available English classifier in any form

Requires English dataset, Multilingual LM & Fine-tuning

(a)

Input in Ukrainian

Прошу надіслати мені повідомлення.

Translation System Ukrainian->English

Input in English

Please, send me a message.

Classification System for English

Label

(b)

Prompt:

This is detecting text formality task. Classify the text into two categories: formal or informal.

Examples: [*corresponding examples*]

Text: [*text in Ukrainian*]
Sentiment:

Label

(c)

Dataset in English

Translation System English->Ukrainian

Dataset in Ukrainian

Dataset in Ukrainian

Monolingual or Multilingual LM

Classification System for Ukrainian

Input in Ukrainian

Classification System for Ukrainian

Label

(d)

Adapter Task

Adapter English

Multilingual LM

Dataset in English

Adapter Task

Adapter English

Multilingual LM

Input in Ukrainian

Adapter Task

Adapter Ukrainian

Multilingual LM

Label

AdapterHub

Question

How to choose a translation system?

Candidates

How to choose a translation system?

OPUS-NLP

Helsinki

open-source

No Language Left Behind

NLLB @ Meta

open-source

Commercial APIs

DeepL

commercial

Results

How to choose a translation system?

OPUS-NLP

- Toxicity translation

NLLB @ Meta

- Neutral sentences translation

DeepL

- Was good in all three tasks

Data sources

Original English (translated sets)

- **Toxicity:** Jigsaw Toxic Comment Classification Challenge
- **Formality:** GYAFC (Rao & Tetreault, 2018)
- **NLI:** SNLI (Bowman et al., 2015)

Original Ukranian (semi-natural sets)

- **Toxicity:** filtered tweets by toxic keywords + news & fiction UD Ukrainian IU dataset
- **Formality:** Ukrainian legal acts and tweets
- **NLI:** Ukrainian legal acts · corpus of Ukrainian fiction· manually created

Data

Data statistics

	Toxicity dataset	Formality dataset	NLI dataset
Train	total: 24616 toxic: 12307 non-toxic: 12309	total: 209124 formal: 104562 informal: 104562	total: 549361 neutral: 182762 contradiction: 183185 entailment: 183414
Val	total: 4000 toxic: 2000 non-toxic: 2000	total: 10272 formal: 4605 informal: 5667	total: 9842 neutral: 3235 contradiction: 3278 entailment: 3329
Test	total: 52294 toxic: 5800 non-toxic: 46494	total: 4853 formal: 2103 informal: 2750	total: 9824 neutral: 3219 contradiction: 3237 entailment: 3368
Semi-natural Test	total: 4214 toxic: 2114 non-toxic: 2088	total: 3000 formal: 1500 informal: 1500	total: 901 neutral: 300 contradiction: 300 entailment: 301

Experiments

Question

How to design a prompt for other language rather than English?

Prompting recipe

Design a prompt for non-English languages

Prompt template

```
# Task description — in English
You are a classifier for toxicity detection.
Given a text, output one of {toxic, non-toxic}.

# Few-shot examples — in the target language
Text: З**балась уже ту ініціативу брати, скільки можна?
Label: toxic

# Your text — in target language
Text: дякую за допомогу, дуже приємно
Label: non-toxic
```

Prompting recipe

Design a prompt for non-English languages

Prompt template

```
# Task description – in English
This is Natural language inference (NLI) task. Determine
whether a given hypothesis is contradiction, entailment
or neutral in relation to a given premise. Reply with
only one word: contradiction, neutral or entailment.

# Few-shot examples – in the target language
Premise: Чоловік у чорній сорочки грає в гольф ззовні.
Hypothesis: Чоловік грає на полі гольфу, щоб відпочити.
Label: neutral.
Premise: Чоловік у чорній сорочки грає в гольф ззовні.
Hypothesis: Чоловік у чорній сорочки обмінюється картами з
дівчиною.
Label: contradiction.
Premise: Чоловік у чорній сорочки грає в гольф ззовні.
Hypothesis: Чоловік у чорній сорочки грає в гольф.
Label: entailment

# Your text – in target language
Text: дякую за допомогу, дуже приємно
Label: non-toxic
```

Experiments

Prompting LLMs

At that point, Mistral was a winner.

	Acc	Pr	Re	F1	Acc	Pr	Re	F1
	Translated Test Set				Semi-natural Test Set			
Toxicity Classification								
LLaMa-2 Prompting	0.51	0.50	0.67	0.42	0.67	0.67	0.49	0.67
LLaMa-3 Prompting	0.61	0.56	0.66	0.55	0.70	0.79	0.67	0.68
Mistral Prompting	**0.86**	**0.68**	**0.74**	**0.70**	**0.76**	**0.81**	**0.76**	**0.75**
FLAN-T5-Backtranslation	—				0.69	0.73	0.69	0.68
Formality Classification								
LLaMa-2 Prompting	0.43	0.22	0.50	0.30	0.50	0.25	0.50	0.33
LLaMa-3 Prompting	0.51	0.45	0.64	0.52	0.78	0.67	0.72	0.71
Mistral Prompting	**0.64**	**0.63**	**0.64**	**0.63**	**0.94**	**0.94**	**0.94**	**0.94**
FLAN-T5-Backtranslation	—				0.62	0.77	0.62	0.56
Natural Language Inference								
LLaMa-2 Prompting	0.36	0.40	0.36	0.34	0.37	0.28	0.36	0.28
LLaMa-3 Prompting	0.55	0.50	**0.57**	0.55	0.66	0.68	0.66	0.66
Mistral Prompting	**0.56**	**0.61**	0.56	**0.56**	**0.71**	**0.72**	**0.69**	**0.69**
FLAN-T5-Backtranslation	—				0.48	0.68	0.49	0.42

Experiments

The whole picture for three tasks

	Acc	Pr	Re	F1	Acc	Pr	Re	F1
	Translated Test Set				Semi-natural Test Set			
Toxicity Classification								
Mistral Prompting	0.86	0.68	0.74	0.70	**0.76**	**0.81**	**0.76**	**0.75**
Backtranslation	—				0.63	0.76	0.56	0.58
Adapter Training	**0.87**	0.66	0.63	0.65	0.58	0.66	0.58	0.52
XLM-R-finetuned	0.81	**0.68**	**0.86**	**0.70**	**0.77**	**0.79**	**0.77**	**0.77**
Formality Classification								
Mistral Prompting	0.64	0.63	0.64	0.63	**0.94**	**0.94**	**0.94**	**0.94**
Backtranslation	—				0.56	0.61	0.39	0.50
Adapter Training	**0.64**	**0.63**	0.63	**0.63**	**0.71**	**0.71**	**0.71**	**0.71**
XLM-R-finetuned	0.57	0.28	0.50	0.36	0.50	0.25	0.50	0.33
Natural Language Inference								
Mistral Prompting	0.56	0.61	0.56	0.56	**0.71**	**0.72**	**0.69**	**0.69**
Backtranslation	—				0.40	0.41	0.63	0.33
Adapter Training	0.44	0.46	0.43	0.41	0.40	0.36	0.40	0.32
XLM-R-finetuned	**0.82**	**0.82**	**0.82**	**0.82**	**0.48**	**0.46**	**0.46**	**0.42**

Experiments

The whole picture for three tasks

For toxicity, fine-tuning of LM is still the best

	Acc	Pr	Re	F1	Acc	Pr	Re	F1
	Translated Test Set				Semi-natural Test Set			
Toxicity Classification								
Mistral Prompting	0.86	0.68	0.74	0.70	**0.76**	**0.81**	**0.76**	**0.75**
Backtranslation	—				0.63	0.76	0.56	0.58
Adapter Training	**0.87**	0.66	0.63	0.65	0.58	0.66	0.58	0.52
XLM-R-finetuned	0.81	**0.68**	**0.86**	**0.70**	**0.77**	**0.79**	**0.77**	**0.77**
Formality Classification								
Mistral Prompting	0.64	0.63	0.64	0.63	**0.94**	**0.94**	**0.94**	**0.94**
Backtranslation	—				0.56	0.61	0.39	0.50
Adapter Training	**0.64**	**0.63**	0.63	**0.63**	**0.71**	**0.71**	**0.71**	**0.71**
XLM-R-finetuned	0.57	0.28	0.50	0.36	0.50	0.25	0.50	0.33
Natural Language Inference								
Mistral Prompting	0.56	0.61	0.56	0.56	**0.71**	**0.72**	**0.69**	**0.69**
Backtranslation	—				0.40	0.41	0.63	0.33
Adapter Training	0.44	0.46	0.43	0.41	0.40	0.36	0.40	0.32
XLM-R-finetuned	**0.82**	**0.82**	**0.82**	**0.82**	**0.48**	**0.46**	**0.46**	**0.42**

Experiments

The whole picture for three tasks

For formality, prompting of LLM is surprisingly very good

	Acc	Pr	Re	F1	Acc	Pr	Re	F1
	Translated Test Set				Semi-natural Test Set			
Toxicity Classification								
Mistral Prompting	0.86	0.68	0.74	0.70	**0.76**	**0.81**	**0.76**	**0.75**
Backtranslation	—				0.63	0.76	0.56	0.58
Adapter Training	**0.87**	0.66	0.63	0.65	0.58	0.66	0.58	0.52
XLM-R-finetuned	0.81	**0.68**	**0.86**	**0.70**	**0.77**	**0.79**	**0.77**	**0.77**
Formality Classification								
Mistral Prompting	0.64	0.63	0.64	0.63	**0.94**	**0.94**	**0.94**	**0.94**
Backtranslation	—				0.56	0.61	0.39	0.50
Adapter Training	**0.64**	**0.63**	0.63	**0.63**	**0.71**	**0.71**	**0.71**	**0.71**
XLM-R-finetuned	0.57	0.28	0.50	0.36	0.50	0.25	0.50	0.33
Natural Language Inference								
Mistral Prompting	0.56	0.61	0.56	0.56	**0.71**	**0.72**	**0.69**	**0.69**
Backtranslation	—				0.40	0.41	0.63	0.33
Adapter Training	0.44	0.46	0.43	0.41	0.40	0.36	0.40	0.32
XLM-R-finetuned	**0.82**	**0.82**	**0.82**	**0.82**	**0.48**	**0.46**	**0.46**	**0.42**

Experiments

The whole picture for three tasks

For NLI, both are good baselines,
but natural data is really needed

	Acc	Pr	Re	F1	Acc	Pr	Re	F1
	Translated Test Set				Semi-natural Test Set			
Toxicity Classification								
Mistral Prompting	0.86	0.68	0.74	0.70	**0.76**	**0.81**	**0.76**	**0.75**
Backtranslation	—				0.63	0.76	0.56	0.58
Adapter Training	**0.87**	0.66	0.63	0.65	0.58	0.66	0.58	0.52
XLM-R-finetuned	0.81	**0.68**	**0.86**	**0.70**	**0.77**	**0.79**	**0.77**	**0.77**
Formality Classification								
Mistral Prompting	0.64	0.63	0.64	0.63	**0.94**	**0.94**	**0.94**	**0.94**
Backtranslation	—				0.56	0.61	0.39	0.50
Adapter Training	**0.64**	**0.63**	0.63	**0.63**	**0.71**	**0.71**	**0.71**	**0.71**
XLM-R-finetuned	0.57	0.28	0.50	0.36	0.50	0.25	0.50	0.33
Natural Language Inference								
Mistral Prompting	0.56	0.61	0.56	0.56	**0.71**	**0.72**	**0.69**	**0.69**
Backtranslation	—				0.40	0.41	0.63	0.33
Adapter Training	0.44	0.46	0.43	0.41	0.40	0.36	0.40	0.32
XLM-R-finetuned	**0.82**	**0.82**	**0.82**	**0.82**	**0.48**	**0.46**	**0.46**	**0.42**

Method	Models	Datasets	Translation Dependence	Data Creation	Fine tuning	# Inference Steps
		Cross-lingual Transfer Methods				
Backtranslation	- Toxicity detection model for the resource-rich language; - Translation model from resource-rich to the target language;	—	✔	✘	✘	3
LLM prompting	- LLM with the knowledge of the resource-rich language and (emerging) knowledge of the target language;	—	✘	✘	✘	1
Adapter Training	- Auto-regressive multilingual LM where the resource-rich and target languages are present; - Language adapter layers for both languages;	- Toxicity classification dataset in the resource-rich language; - Corpus for translation between the resource-rich and target languages;	✘	✘	✔	1
		Data Acquisition Methods				
Training Data Translation	- Translation model to the target language; - Auto-regressive multilingual or monolingual LM for the target language;	- Toxicity classification dataset in the resource-rich language;	✔	✔	✔	1
Semi-synthetic data by keywords filtering	- Embedding model of texts in the target language;	- Texts in the target language; - List of toxic keywords in the target language;	✘	✔	✔	1
Crowdsourcing data filtering	- Embedding model of texts in the target language;	- Texts in the target language;	✘	✔	✔	1

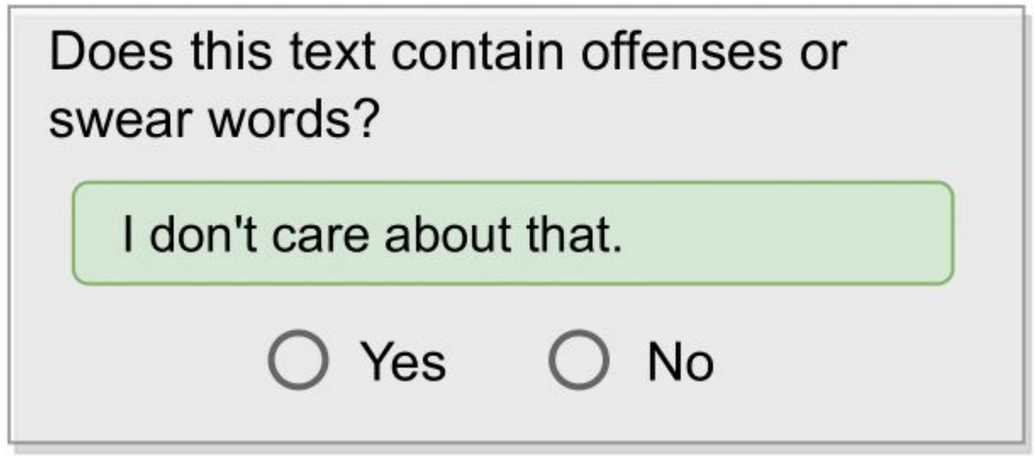
Additional case

Data collection for toxicity classification

Filtering with crowdsourcing to produce a trusted Ukrainian test set.

Pipeline

Filtering with crowdsourcing



Quality control

- Native Ukrainian speakers
- Training · Exam · Control tasks

Published

Toxicity classification in Ukrainian

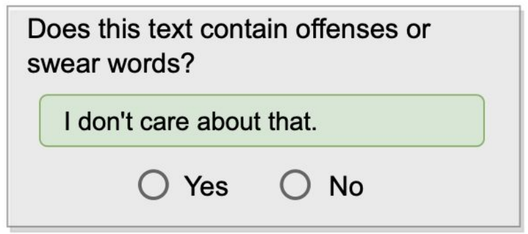


Figure 1: Interface (translated into English for illustration) of the toxicity classification task for data collection with crowdsourcing.

	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
	Translated Test Set			Semi-synthetic Test Set			Crowdsourced Test Set		
Prompting of LLMs									
LLaMa-2 Prompting	0.50	0.67	0.42	0.67	0.49	0.67	0.24	0.50	0.32
Mistral Prompting	**0.68**	**0.74**	**0.70**	**0.81**	**0.76**	**0.75**	**0.56**	**0.68**	**0.52**
Cross-lingual transfer approaches									
Backtranslation		—		**0.76**	**0.56**	**0.58**	**0.75**	**0.68**	**0.65**
Adapter Training	0.66	0.63	0.65	0.66	0.58	0.52	0.64	0.58	0.53
Fine-tuning of LMs on different types of data									
XLM-R-finetuned-translated	0.68	0.86	0.70	0.79	0.77	0.77	0.70	**0.68**	**0.67**
XLM-R-finetuned-semisynthetic	0.59	0.53	0.53	0.99	0.99	0.99	**0.75**	0.57	0.48
XLM-R-finetuned-crowdsourced	**0.61**	**0.63**	**0.62**	**0.93**	**0.93**	**0.93**	0.99	0.99	0.99

Even 5k of original-language and culturally-specific data brings significant improvements!

Dementieva, Khylenko, Babakov, Groh · Toxicity Classification in Ukrainian · WOAH, NAACL 2024
Cross-lingual Text Classification Transfer: The Case of Ukrainian COLING 2025

Main takeaways

- **It is possible** to create cross-lingual classifiers
- LLMs can be **extremely bad** in not so resource-rich languages.
  Mistral showed the most **adequate** performance on non-English data.
- **Fine-tuning of XLM-R** even on the **translated data** can be a strong baseline.
- **Native speakers** are needed to check the quality of **translation**;
  natural test set would be highly beneficial for the evaluation.

Q&A

Questions?

Benchmark

EMOBENCH-UA

A benchmark dataset for emotion detection in Ukrainian.

EMOBENCH-UA

Main goal

Deliver the **first reliable Ukrainian-native** benchmark for fine-grained emotion detection — and measure how current models fare on it.

☺ Six emotions

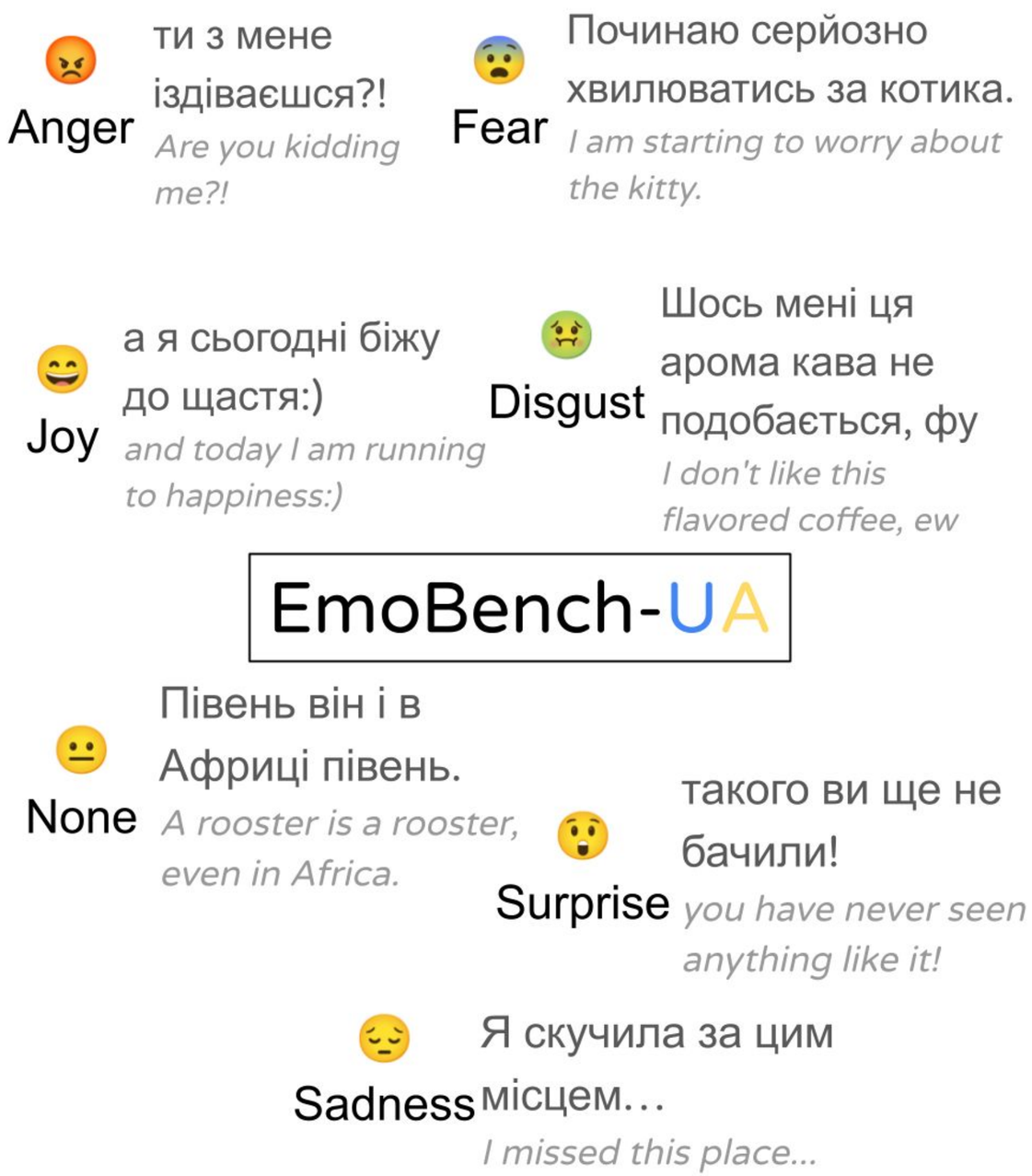
love · anger · fear · sadness · surprise · joy + neutral

Native texts

Tweets, fiction, news — originally written in Ukrainian.

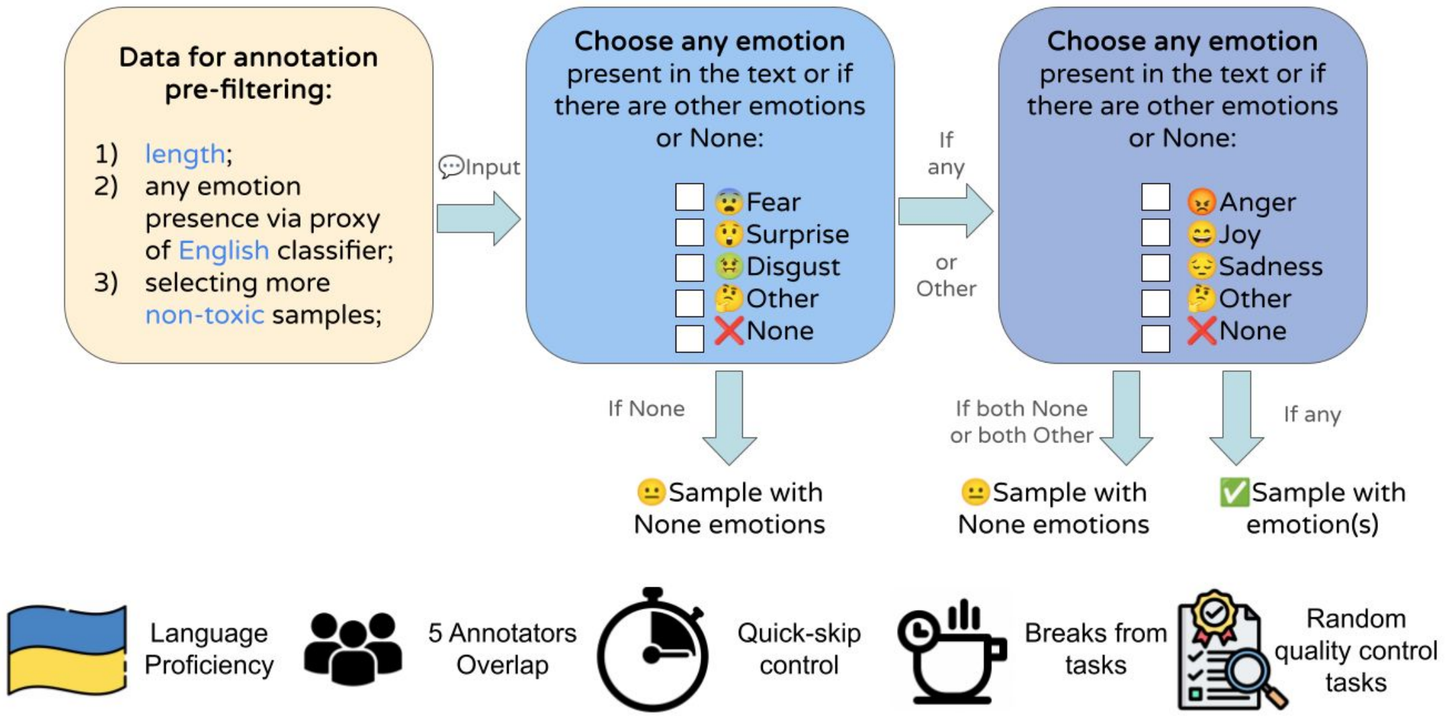
Rigorous QC

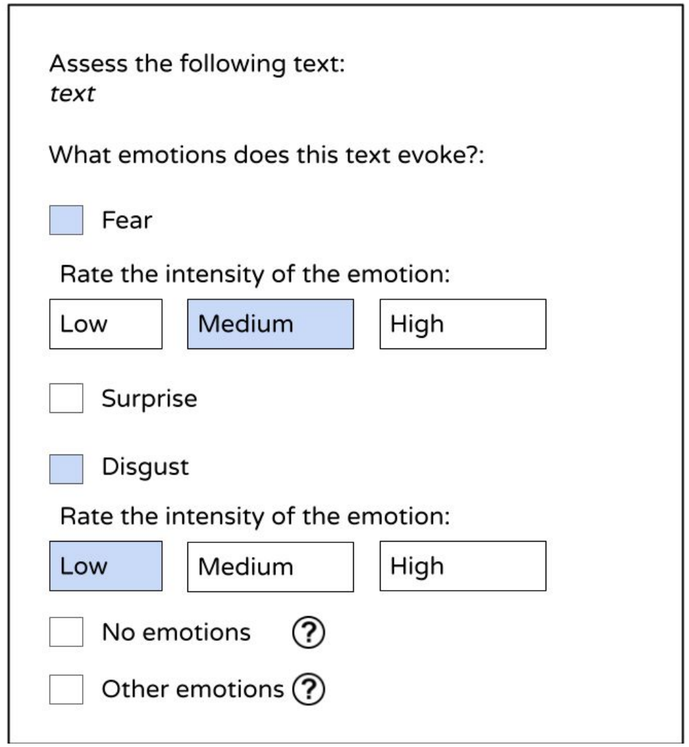
Overlap 5, control tasks, expert audit.



Pipeline

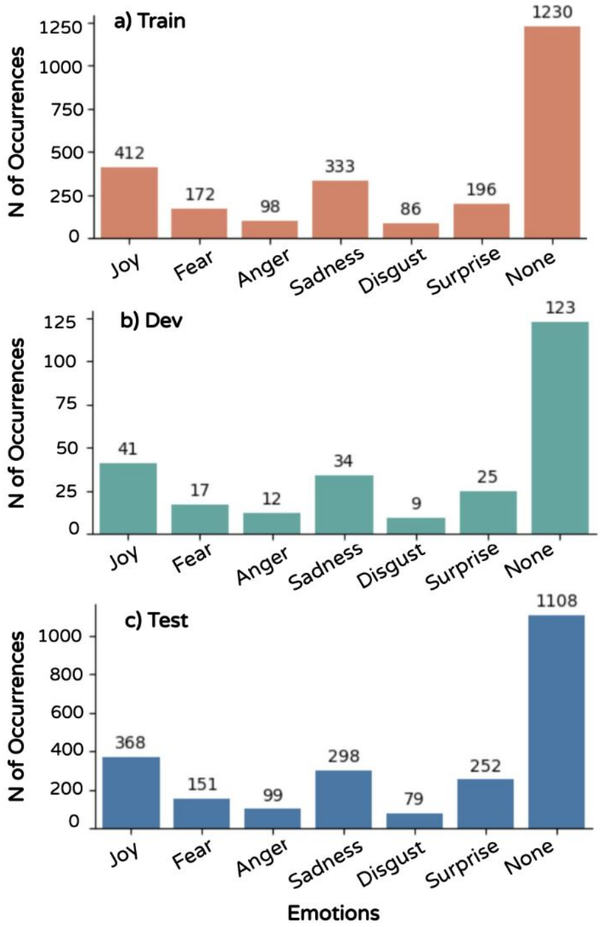
Data collection pipeline





Final data

Final data statistics



(a) Distribution of Labels

Joy					Fear				
щастя *happiness*	вітаю *congratulations*	дякую *thank you*	народження *birth*	святом *holiday*	боїться *afraid*	страшно *scared*	злякати *scare*	лякає *scares*	серце *heart*
настрій *mood*	приємно *nice*	вітання *congrats*	життя *life*	гарно *beautiful*	переживати *worry*	здається *seems*	найбільше *most of all*	жах *horror*	мама *mom*
Anger					**Sadness**				
бісить *pisses off*	гірше *worse*	злий *angry*	клятий *damn*	влада *government*	сумно *sad*	гірше *worse*	шкода *pity*	скучити *miss*	люди *people*
вибішує *pisses off*	ненавидіти *hate*	боже *oh god*	ти *you*	день *day*	печаль *sadness*	нема *no*	хочеться *want*	жаль *regret*	сум *sorrow*
Disgust					**Surprise**				
гірше *worse*	фу *ew*	запах *smell*	лайно *shit*	пахнути *smell*	дивно *weird*	думати *think*	розуміти *understand*	очікувати *expect*	боже *oh god*
гидко *disgusting*	прямо *straight*	дихати *breathe*	їсти *eat*	гнилий *rotten*	дивний *weird*	серйозно *seriously*	реакція *reaction*	чудовий *amazing*	нащо *why*

None

хотіти *want*	вчора *yesterday*	спати *sleep*	день *day*	робота *job*
вночі *at night*	знати *know*	бачити *see*	вдома *at home*	завтра *tomorrow*

(b) Keywords per Emotion

Figure 4: EMOBENCH-UA statistics per sets and emotions.

Setups

Experimental setups

- Keywords Based baselines vs Tuned encoders
- Ukrainian Monolingual vs Multilingual encoders
- Tuned encoders vs LLMs
- Prompting in English vs Ukrainian

Results

	Joy	Fear	Anger	Sadness	Disgust	Surprise	None	Pr	Re	F1
					Linguistic-based Approaches					
Keywords	0.30	0.15	0.08	0.21	0.10	0.15	0.25	0.24	0.24	0.22
Logistic Regression	**0.64**	**0.72**	**0.49**	**0.59**	**0.49**	**0.61**	0.67	0.51	**0.22**	**0.29**
Random Forest	0.61	0.69	**0.49**	**0.59**	**0.49**	0.60	**0.68**	**0.58**	0.21	0.27
					Translation to English					
DistillRoBERTa-Emo-EN	0.56	0.55	0.31	0.52	0.23	0.47	0.55	0.40	0.61	0.45
					Transformer-based Encoders					
LaBSe	0.67	0.73	0.30	0.65	0.33	0.54	0.80	0.57	0.59	0.57
Geotrend-BERT	**0.58**	**0.59**	**0.08**	**0.50**	**0.11**	**0.40**	**0.73**	**0.46**	**0.43**	**0.43**
mBERT	0.46	0.24	0.01	0.45	0.02	0.33	0.73	0.33	0.33	0.32
UKR-RoBERTa Base	0.65	0.58	0.14	0.50	**0.21**	0.49	0.74	0.51	0.45	0.47
XLM-RoBERTa Base	0.61	0.31	0.00	0.33	0.01	0.19	0.75	0.33	0.31	0.31
XLM-RoBERTa Large	**0.73**	**0.79**	**0.20**	**0.68**	0.00	**0.60**	**0.80**	0.52	**0.58**	**0.54**
Twitter-XLM-RoBERTa	0.72	0.76	0.13	0.64	0.07	0.54	0.79	**0.66**	0.51	0.52
Glot500 Base	0.01	0.02	0.03	0.18	0.00	0.01	0.64	0.24	0.19	0.13
Multilingual-E5 Base	0.71	0.73	0.01	0.52	0.00	0.50	0.77	0.49	0.45	0.46
Multilingual-E5 Large	**0.73**	**0.81**	**0.31**	**0.69**	**0.35**	**0.60**	**0.81**	**0.65**	**0.62**	**0.62**
					LLMs Prompting					
EuroLLM-1.7B (ENG)	**0.46**	**0.31**	**0.15**	**0.37**	**0.18**	0.09	**0.28**	**0.26**	**0.38**	**0.26**
EuroLLM-1.7B (UKR)	0.38	0.30	0.11	0.27	0.10	**0.11**	0.25	0.25	0.24	0.22
Spivavtor-XXL (ENG)	**0.39**	0.03	**0.15**	**0.13**	0.00	0.01	**0.69**	**0.68**	0.20	**0.20**
Spivavtor-XXL (UKR)	0.32	**0.08**	0.14	**0.13**	**0.08**	**0.13**	0.29	0.17	**0.28**	0.17
MamayLM-9B (ENG)	**0.63**	**0.62**	**0.54**	**0.64**	0.38	**0.31**	**0.67**	**0.46**	**0.73**	**0.54**
MamayLM-9B (UKR)	0.61	0.61	0.47	0.52	**0.47**	0.24	0.41	0.44	**0.73**	0.48
Mistral-7B (ENG)	0.52	**0.58**	**0.33**	**0.49**	**0.32**	**0.37**	**0.52**	**0.37**	**0.73**	**0.45**
Mistral-7B (UKR)	**0.55**	0.37	0.28	0.47	0.19	0.24	0.33	0.32	0.71	0.35
Mixtral-8x7B (ENG)	**0.49**	**0.37**	**0.34**	**0.51**	**0.25**	**0.25**	0.66	**0.32**	**0.74**	**0.41**
Mixtral-8x7B (UKR)	0.48	0.35	0.19	0.47	0.21	0.22	**0.71**	0.27	0.73	0.37
LLaMA 3 8B (ENG)	**0.56**	0.65	**0.36**	**0.54**	**0.29**	**0.25**	**0.39**	**0.43**	**0.56**	**0.43**
LLaMA 3 8B (UKR)	0.30	**0.67**	0.29	0.45	0.15	**0.25**	0.10	0.38	0.53	0.31
LLaMA 3.3 70B (ENG)	**0.64**	0.63	**0.47**	0.62	**0.26**	0.32	**0.43**	0.44	**0.79**	**0.48**
LLaMA 3.3 70B (UKR)	0.58	**0.68**	0.34	**0.71**	0.18	**0.33**	0.36	**0.45**	0.64	0.46
Qwen3-4B (ENG)	**0.65**	**0.66**	0.39	**0.56**	**0.34**	**0.35**	**0.52**	**0.45**	**0.72**	**0.49**
Qwen3-4B (UKR)	0.63	0.62	**0.42**	0.54	0.18	0.34	0.33	0.43	0.69	0.44
DeepSeek-R1-Qwen (ENG)	0.63	0.61	**0.43**	**0.64**	**0.45**	**0.46**	0.60	**0.48**	**0.75**	**0.55**
DeepSeek-R1-Qwen (UKR)	**0.68**	**0.66**	0.40	0.57	0.29	0.38	**0.68**	0.46	0.66	0.52
DeepSeek-R1-LLaMA (ENG)	**0.67**	**0.69**	**0.49**	**0.71**	**0.52**	0.47	0.67	**0.54**	**0.72**	**0.60**
DeepSeek-R1-LLaMA (UKR)	**0.67**	0.64	0.45	0.69	0.33	**0.51**	**0.69**	0.51	0.69	0.57
DeepSeek-V3 (ENG)	**0.73**	**0.74**	0.60	**0.72**	**0.57**	0.41	**0.78**	**0.60**	0.72	**0.65**
DeepSeek-V3 (UKR)	0.71	0.66	**0.61**	**0.72**	0.48	**0.42**	0.71	0.54	**0.81**	0.62

Results

Surprise, surprise:
Even LR can achieve quite high results for some emotions.

	Joy	Fear	Anger	Sadness	Disgust	Surprise	None	Pr	Re	F1
					Linguistic-based Approaches					
Keywords	0.30	0.15	0.08	0.21	0.10	0.15	0.25	0.24	0.24	0.22
Logistic Regression	**0.64**	**0.72**	**0.49**	**0.59**	**0.49**	**0.61**	0.67	0.51	**0.22**	**0.29**
Random Forest	0.61	0.69	**0.49**	**0.59**	**0.49**	0.60	**0.68**	**0.58**	0.21	0.27
					Translation to English					
DistillRoBERTa-Emo-EN	0.56	0.55	0.31	0.52	0.23	0.47	0.55	0.40	0.61	0.45
					Transformer-based Encoders					
LaBSe	0.67	0.73	0.30	0.65	0.33	0.54	0.80	0.57	0.59	0.57
Geotrend-BERT	**0.58**	**0.59**	**0.08**	**0.50**	**0.11**	**0.40**	**0.73**	**0.46**	**0.43**	**0.43**
mBERT	0.46	0.24	0.01	0.45	0.02	0.33	0.73	0.33	0.33	0.32
UKR-RoBERTa Base	0.65	0.58	0.14	0.50	**0.21**	0.49	0.74	0.51	0.45	0.47
XLM-RoBERTa Base	0.61	0.31	0.00	0.33	0.01	0.19	0.75	0.33	0.31	0.31
XLM-RoBERTa Large	**0.73**	**0.79**	**0.20**	**0.68**	0.00	**0.60**	**0.80**	0.52	**0.58**	**0.54**
Twitter-XLM-RoBERTa	0.72	0.76	0.13	0.64	0.07	0.54	0.79	**0.66**	0.51	0.52
Glot500 Base	0.01	0.02	0.03	0.18	0.00	0.01	0.64	0.24	0.19	0.13
Multilingual-E5 Base	0.71	0.73	0.01	0.52	0.00	0.50	0.77	0.49	0.45	0.46
Multilingual-E5 Large	**0.73**	**0.81**	**0.31**	**0.69**	**0.35**	**0.60**	**0.81**	**0.65**	**0.62**	**0.62**
					LLMs Prompting					
EuroLLM-1.7B (ENG)	**0.46**	**0.31**	**0.15**	**0.37**	**0.18**	0.09	**0.28**	**0.26**	**0.38**	**0.26**
EuroLLM-1.7B (UKR)	0.38	0.30	0.11	0.27	0.10	**0.11**	0.25	0.25	0.24	0.22
Spivavtor-XXL (ENG)	**0.39**	0.03	**0.15**	**0.13**	0.00	0.01	**0.69**	**0.68**	0.20	**0.20**
Spivavtor-XXL (UKR)	0.32	**0.08**	0.14	**0.13**	**0.08**	**0.13**	0.29	0.17	**0.28**	0.17
MamayLM-9B (ENG)	**0.63**	**0.62**	**0.54**	**0.64**	0.38	**0.31**	**0.67**	**0.46**	**0.73**	**0.54**
MamayLM-9B (UKR)	0.61	0.61	0.47	0.52	**0.47**	0.24	0.41	0.44	**0.73**	0.48
Mistral-7B (ENG)	0.52	**0.58**	**0.33**	**0.49**	**0.32**	**0.37**	**0.52**	**0.37**	**0.73**	**0.45**
Mistral-7B (UKR)	**0.55**	0.37	0.28	0.47	0.19	0.24	0.33	0.32	0.71	0.35
Mixtral-8x7B (ENG)	**0.49**	**0.37**	**0.34**	**0.51**	**0.25**	**0.25**	0.66	**0.32**	**0.74**	**0.41**
Mixtral-8x7B (UKR)	0.48	0.35	0.19	0.47	0.21	0.22	**0.71**	0.27	0.73	0.37
LLaMA 3 8B (ENG)	**0.56**	0.65	**0.36**	**0.54**	**0.29**	**0.25**	**0.39**	**0.43**	**0.56**	**0.43**
LLaMA 3 8B (UKR)	0.30	**0.67**	0.29	0.45	0.15	**0.25**	0.10	0.38	0.53	0.31
LLaMA 3.3 70B (ENG)	**0.64**	0.63	**0.47**	0.62	**0.26**	0.32	**0.43**	0.44	**0.79**	**0.48**
LLaMA 3.3 70B (UKR)	0.58	**0.68**	0.34	**0.71**	0.18	**0.33**	0.36	**0.45**	0.64	0.46
Qwen3-4B (ENG)	**0.65**	**0.66**	0.39	**0.56**	**0.34**	**0.35**	**0.52**	**0.45**	**0.72**	**0.49**
Qwen3-4B (UKR)	0.63	0.62	**0.42**	0.54	0.18	0.34	0.33	0.43	0.69	0.44
DeepSeek-R1-Qwen (ENG)	0.63	0.61	**0.43**	**0.64**	**0.45**	**0.46**	0.60	**0.48**	**0.75**	**0.55**
DeepSeek-R1-Qwen (UKR)	**0.68**	**0.66**	0.40	0.57	0.29	0.38	**0.68**	0.46	0.66	0.52
DeepSeek-R1-LLaMA (ENG)	**0.67**	**0.69**	**0.49**	**0.71**	**0.52**	0.47	0.67	**0.54**	**0.72**	**0.60**
DeepSeek-R1-LLaMA (UKR)	**0.67**	0.64	0.45	0.69	0.33	**0.51**	**0.69**	0.51	0.69	0.57
DeepSeek-V3 (ENG)	**0.73**	**0.74**	0.60	**0.72**	**0.57**	0.41	**0.78**	**0.60**	0.72	**0.65**
DeepSeek-V3 (UKR)	0.71	0.66	**0.61**	**0.72**	0.48	**0.42**	0.71	0.54	**0.81**	0.62

Results

Sad fact: unfortunately, specifically Ukrainian-BERT models did not so good results.
Multilingual ones — the bigger the better.

	Joy	Fear	Anger	Sadness	Disgust	Surprise	None	Pr	Re	F1
					Linguistic-based Approaches					
Keywords	0.30	0.15	0.08	0.21	0.10	0.15	0.25	0.24	0.24	0.22
Logistic Regression	**0.64**	**0.72**	**0.49**	**0.59**	**0.49**	**0.61**	0.67	0.51	**0.22**	**0.29**
Random Forest	0.61	0.69	**0.49**	**0.59**	**0.49**	0.60	**0.68**	**0.58**	0.21	0.27
					Translation to English					
DistillRoBERTa-Emo-EN	0.56	0.55	0.31	0.52	0.23	0.47	0.55	0.40	0.61	0.45
					Transformer-based Encoders					
LaBSe	0.67	0.73	0.30	0.65	0.33	0.54	0.80	0.57	0.59	0.57
Geotrend-BERT	**0.58**	**0.59**	**0.08**	**0.50**	**0.11**	**0.40**	**0.73**	**0.46**	**0.43**	**0.43**
mBERT	0.46	0.24	0.01	0.45	0.02	0.33	0.73	0.33	0.33	0.32
UKR-RoBERTa Base	0.65	0.58	0.14	0.50	**0.21**	0.49	0.74	0.51	0.45	0.47
XLM-RoBERTa Base	0.61	0.31	0.00	0.33	0.01	0.19	0.75	0.33	0.31	0.31
XLM-RoBERTa Large	**0.73**	**0.79**	**0.20**	**0.68**	0.00	**0.60**	**0.80**	0.52	**0.58**	**0.54**
Twitter-XLM-RoBERTa	0.72	0.76	0.13	0.64	0.07	0.54	0.79	**0.66**	0.51	0.52
Glot500 Base	0.01	0.02	0.03	0.18	0.00	0.01	0.64	0.24	0.19	0.13
Multilingual-E5 Base	0.71	0.73	0.01	0.52	0.00	0.50	0.77	0.49	0.45	0.46
Multilingual-E5 Large	**0.73**	**0.81**	**0.31**	**0.69**	**0.35**	**0.60**	**0.81**	**0.65**	**0.62**	**0.62**
					LLMs Prompting					
EuroLLM-1.7B (ENG)	**0.46**	**0.31**	**0.15**	**0.37**	**0.18**	0.09	**0.28**	**0.26**	**0.38**	**0.26**
EuroLLM-1.7B (UKR)	0.38	0.30	0.11	0.27	0.10	**0.11**	0.25	0.25	0.24	0.22
Spivavtor-XXL (ENG)	**0.39**	0.03	**0.15**	**0.13**	0.00	0.01	**0.69**	**0.68**	0.20	**0.20**
Spivavtor-XXL (UKR)	0.32	**0.08**	0.14	**0.13**	**0.08**	**0.13**	0.29	0.17	**0.28**	0.17
MamayLM-9B (ENG)	**0.63**	**0.62**	**0.54**	**0.64**	0.38	**0.31**	**0.67**	**0.46**	**0.73**	**0.54**
MamayLM-9B (UKR)	0.61	0.61	0.47	0.52	**0.47**	0.24	0.41	0.44	**0.73**	0.48
Mistral-7B (ENG)	0.52	**0.58**	**0.33**	**0.49**	**0.32**	**0.37**	**0.52**	**0.37**	**0.73**	**0.45**
Mistral-7B (UKR)	**0.55**	0.37	0.28	0.47	0.19	0.24	0.33	0.32	0.71	0.35
Mixtral-8x7B (ENG)	**0.49**	**0.37**	**0.34**	**0.51**	**0.25**	**0.25**	0.66	**0.32**	**0.74**	**0.41**
Mixtral-8x7B (UKR)	0.48	0.35	0.19	0.47	0.21	0.22	**0.71**	0.27	0.73	0.37
LLaMA 3 8B (ENG)	**0.56**	0.65	**0.36**	**0.54**	**0.29**	**0.25**	**0.39**	**0.43**	**0.56**	**0.43**
LLaMA 3 8B (UKR)	0.30	**0.67**	0.29	0.45	0.15	**0.25**	0.10	0.38	0.53	0.31
LLaMA 3.3 70B (ENG)	**0.64**	0.63	**0.47**	0.62	**0.26**	0.32	**0.43**	0.44	**0.79**	**0.48**
LLaMA 3.3 70B (UKR)	0.58	**0.68**	0.34	**0.71**	0.18	**0.33**	0.36	**0.45**	0.64	0.46
Qwen3-4B (ENG)	**0.65**	**0.66**	0.39	**0.56**	**0.34**	**0.35**	**0.52**	**0.45**	**0.72**	**0.49**
Qwen3-4B (UKR)	0.63	0.62	**0.42**	0.54	0.18	0.34	0.33	0.43	0.69	0.44
DeepSeek-R1-Qwen (ENG)	0.63	0.61	**0.43**	**0.64**	**0.45**	**0.46**	0.60	**0.48**	**0.75**	**0.55**
DeepSeek-R1-Qwen (UKR)	**0.68**	**0.66**	0.40	0.57	0.29	0.38	**0.68**	0.46	0.66	0.52
DeepSeek-R1-LLaMA (ENG)	**0.67**	**0.69**	**0.49**	**0.71**	**0.52**	0.47	0.67	**0.54**	**0.72**	**0.60**
DeepSeek-R1-LLaMA (UKR)	**0.67**	0.64	0.45	0.69	0.33	**0.51**	**0.69**	0.51	0.69	0.57
DeepSeek-V3 (ENG)	**0.73**	**0.74**	0.60	**0.72**	**0.57**	0.41	**0.78**	**0.60**	0.72	**0.65**
DeepSeek-V3 (UKR)	0.71	0.66	**0.61**	**0.72**	0.48	**0.42**	0.71	0.54	**0.81**	0.62

Results

This might make you angry, but still for every LLM prompting in English was a better strategy.

	Joy	Fear	Anger	Sadness	Disgust	Surprise	None	Pr	Re	F1
					Linguistic-based Approaches					
Keywords	0.30	0.15	0.08	0.21	0.10	0.15	0.25	0.24	0.24	0.22
Logistic Regression	**0.64**	**0.72**	**0.49**	**0.59**	**0.49**	**0.61**	0.67	0.51	**0.22**	**0.29**
Random Forest	0.61	0.69	**0.49**	**0.59**	**0.49**	0.60	**0.68**	**0.58**	0.21	0.27
					Translation to English					
DistillRoBERTa-Emo-EN	0.56	0.55	0.31	0.52	0.23	0.47	0.55	0.40	0.61	0.45
					Transformer-based Encoders					
LaBSe	0.67	0.73	0.30	0.65	0.33	0.54	0.80	0.57	0.59	0.57
Geotrend-BERT	**0.58**	**0.59**	**0.08**	**0.50**	**0.11**	**0.40**	**0.73**	**0.46**	**0.43**	**0.43**
mBERT	0.46	0.24	0.01	0.45	0.02	0.33	0.73	0.33	0.33	0.32
UKR-RoBERTa Base	0.65	0.58	0.14	0.50	**0.21**	0.49	0.74	0.51	0.45	0.47
XLM-RoBERTa Base	0.61	0.31	0.00	0.33	0.01	0.19	0.75	0.33	0.31	0.31
XLM-RoBERTa Large	**0.73**	**0.79**	**0.20**	**0.68**	0.00	**0.60**	**0.80**	0.52	**0.58**	**0.54**
Twitter-XLM-RoBERTa	0.72	0.76	0.13	0.64	0.07	0.54	0.79	**0.66**	0.51	0.52
Glot500 Base	0.01	0.02	0.03	0.18	0.00	0.01	0.64	0.24	0.19	0.13
Multilingual-E5 Base	0.71	0.73	0.01	0.52	0.00	0.50	0.77	0.49	0.45	0.46
Multilingual-E5 Large	**0.73**	**0.81**	**0.31**	**0.69**	**0.35**	**0.60**	**0.81**	**0.65**	**0.62**	**0.62**
					LLMs Prompting					
EuroLLM-1.7B (ENG)	**0.46**	**0.31**	**0.15**	**0.37**	**0.18**	0.09	**0.28**	**0.26**	**0.38**	**0.26**
EuroLLM-1.7B (UKR)	0.38	0.30	0.11	0.27	0.10	**0.11**	0.25	0.25	0.24	0.22
Spivavtor-XXL (ENG)	**0.39**	0.03	**0.15**	**0.13**	0.00	0.01	**0.69**	**0.68**	0.20	**0.20**
Spivavtor-XXL (UKR)	0.32	**0.08**	0.14	**0.13**	**0.08**	**0.13**	0.29	0.17	**0.28**	0.17
MamayLM-9B (ENG)	**0.63**	**0.62**	**0.54**	**0.64**	0.38	**0.31**	**0.67**	**0.46**	**0.73**	**0.54**
MamayLM-9B (UKR)	0.61	0.61	0.47	0.52	**0.47**	0.24	0.41	0.44	**0.73**	0.48
Mistral-7B (ENG)	0.52	**0.58**	**0.33**	**0.49**	**0.32**	**0.37**	**0.52**	**0.37**	**0.73**	**0.45**
Mistral-7B (UKR)	**0.55**	0.37	0.28	0.47	0.19	0.24	0.33	0.32	0.71	0.35
Mixtral-8x7B (ENG)	**0.49**	**0.37**	**0.34**	**0.51**	**0.25**	**0.25**	0.66	**0.32**	**0.74**	**0.41**
Mixtral-8x7B (UKR)	0.48	0.35	0.19	0.47	0.21	0.22	**0.71**	0.27	0.73	0.37
LLaMA 3 8B (ENG)	**0.56**	0.65	**0.36**	**0.54**	**0.29**	**0.25**	**0.39**	**0.43**	**0.56**	**0.43**
LLaMA 3 8B (UKR)	0.30	**0.67**	0.29	0.45	0.15	**0.25**	0.10	0.38	0.53	0.31
LLaMA 3.3 70B (ENG)	**0.64**	0.63	**0.47**	0.62	**0.26**	0.32	**0.43**	0.44	**0.79**	**0.48**
LLaMA 3.3 70B (UKR)	0.58	**0.68**	0.34	**0.71**	0.18	**0.33**	0.36	**0.45**	0.64	0.46
Qwen3-4B (ENG)	**0.65**	**0.66**	0.39	**0.56**	**0.34**	**0.35**	**0.52**	**0.45**	**0.72**	**0.49**
Qwen3-4B (UKR)	0.63	0.62	**0.42**	0.54	0.18	0.34	0.33	0.43	0.69	0.44
DeepSeek-R1-Qwen (ENG)	0.63	0.61	**0.43**	**0.64**	**0.45**	**0.46**	0.60	**0.48**	**0.75**	**0.55**
DeepSeek-R1-Qwen (UKR)	**0.68**	**0.66**	0.40	0.57	0.29	0.38	**0.68**	0.46	0.66	0.52
DeepSeek-R1-LLaMA (ENG)	**0.67**	**0.69**	**0.49**	**0.71**	**0.52**	0.47	0.67	**0.54**	**0.72**	**0.60**
DeepSeek-R1-LLaMA (UKR)	**0.67**	0.64	0.45	0.69	0.33	**0.51**	**0.69**	0.51	0.69	0.57
DeepSeek-V3 (ENG)	**0.73**	**0.74**	0.60	**0.72**	**0.57**	0.41	**0.78**	**0.60**	0.72	**0.65**
DeepSeek-V3 (UKR)	0.71	0.66	**0.61**	**0.72**	0.48	**0.42**	0.71	0.54	**0.81**	0.62

Results

Results

In the end, what brings fun and joy is the fact that DeepSeek–even if it was not explicitly train in Ukrainian–achieved the top performance.

	Joy	Fear	Anger	Sadness	Disgust	Surprise	None	Pr	Re	F1
					Linguistic-based Approaches					
Keywords	0.30	0.15	0.08	0.21	0.10	0.15	0.25	0.24	0.24	0.22
Logistic Regression	**0.64**	**0.72**	**0.49**	**0.59**	**0.49**	**0.61**	0.67	0.51	**0.22**	**0.29**
Random Forest	0.61	0.69	**0.49**	**0.59**	**0.49**	0.60	**0.68**	**0.58**	0.21	0.27
					Translation to English					
DistillRoBERTa-Emo-EN	0.56	0.55	0.31	0.52	0.23	0.47	0.55	0.40	0.61	0.45
					Transformer-based Encoders					
LaBSe	0.67	0.73	0.30	0.65	0.33	0.54	0.80	0.57	0.59	0.57
Geotrend-BERT	**0.58**	**0.59**	**0.08**	**0.50**	**0.11**	**0.40**	**0.73**	**0.46**	**0.43**	**0.43**
mBERT	0.46	0.24	0.01	0.45	0.02	0.33	0.73	0.33	0.33	0.32
UKR-RoBERTa Base	0.65	0.58	0.14	0.50	**0.21**	0.49	0.74	0.51	0.45	0.47
XLM-RoBERTa Base	0.61	0.31	0.00	0.33	0.01	0.19	0.75	0.33	0.31	0.31
XLM-RoBERTa Large	**0.73**	**0.79**	**0.20**	**0.68**	0.00	**0.60**	**0.80**	0.52	**0.58**	**0.54**
Twitter-XLM-RoBERTa	0.72	0.76	0.13	0.64	0.07	0.54	0.79	**0.66**	0.51	0.52
Glot500 Base	0.01	0.02	0.03	0.18	0.00	0.01	0.64	0.24	0.19	0.13
Multilingual-E5 Base	0.71	0.73	0.01	0.52	0.00	0.50	0.77	0.49	0.45	0.46
Multilingual-E5 Large	**0.73**	**0.81**	**0.31**	**0.69**	**0.35**	**0.60**	**0.81**	**0.65**	**0.62**	**0.62**
					LLMs Prompting					
EuroLLM-1.7B (ENG)	**0.46**	**0.31**	**0.15**	**0.37**	**0.18**	0.09	**0.28**	**0.26**	**0.38**	**0.26**
EuroLLM-1.7B (UKR)	0.38	0.30	0.11	0.27	0.10	**0.11**	0.25	0.25	0.24	0.22
Spivavtor-XXL (ENG)	**0.39**	0.03	**0.15**	**0.13**	0.00	0.01	**0.69**	**0.68**	0.20	**0.20**
Spivavtor-XXL (UKR)	0.32	**0.08**	0.14	**0.13**	**0.08**	**0.13**	0.29	0.17	**0.28**	0.17
MamayLM-9B (ENG)	**0.63**	**0.62**	**0.54**	**0.64**	0.38	**0.31**	**0.67**	**0.46**	**0.73**	**0.54**
MamayLM-9B (UKR)	0.61	0.61	0.47	0.52	**0.47**	0.24	0.41	0.44	**0.73**	0.48
Mistral-7B (ENG)	0.52	**0.58**	**0.33**	**0.49**	**0.32**	**0.37**	**0.52**	**0.37**	**0.73**	**0.45**
Mistral-7B (UKR)	**0.55**	0.37	0.28	0.47	0.19	0.24	0.33	0.32	0.71	0.35
Mixtral-8x7B (ENG)	**0.49**	**0.37**	**0.34**	**0.51**	**0.25**	**0.25**	0.66	**0.32**	**0.74**	**0.41**
Mixtral-8x7B (UKR)	0.48	0.35	0.19	0.47	0.21	0.22	**0.71**	0.27	0.73	0.37
LLaMA 3 8B (ENG)	**0.56**	0.65	**0.36**	**0.54**	**0.29**	**0.25**	**0.39**	**0.43**	**0.56**	**0.43**
LLaMA 3 8B (UKR)	0.30	**0.67**	0.29	0.45	0.15	**0.25**	0.10	0.38	0.53	0.31
LLaMA 3.3 70B (ENG)	**0.64**	0.63	**0.47**	0.62	**0.26**	0.32	**0.43**	0.44	**0.79**	**0.48**
LLaMA 3.3 70B (UKR)	0.58	**0.68**	0.34	**0.71**	0.18	**0.33**	0.36	**0.45**	0.64	0.46
Qwen3-4B (ENG)	**0.65**	**0.66**	0.39	**0.56**	**0.34**	**0.35**	**0.52**	**0.45**	**0.72**	**0.49**
Qwen3-4B (UKR)	0.63	0.62	**0.42**	0.54	0.18	0.34	0.33	0.43	0.69	0.44
DeepSeek-R1-Qwen (ENG)	0.63	0.61	**0.43**	**0.64**	**0.45**	**0.46**	0.60	**0.48**	**0.75**	**0.55**
DeepSeek-R1-Qwen (UKR)	**0.68**	**0.66**	0.40	0.57	0.29	0.38	**0.68**	0.46	0.66	0.52
DeepSeek-R1-LLaMA (ENG)	**0.67**	**0.69**	**0.49**	**0.71**	**0.52**	0.47	0.67	**0.54**	**0.72**	**0.60**
DeepSeek-R1-LLaMA (UKR)	**0.67**	0.64	0.45	0.69	0.33	**0.51**	**0.69**	0.51	0.69	0.57
DeepSeek-V3 (ENG)	**0.73**	**0.74**	0.60	**0.72**	**0.57**	0.41	**0.78**	**0.60**	0.72	**0.65**
DeepSeek-V3 (UKR)	0.71	0.66	**0.61**	**0.72**	0.48	**0.42**	0.71	0.54	**0.81**	0.62

Results

Using Translation from English as a Proxy

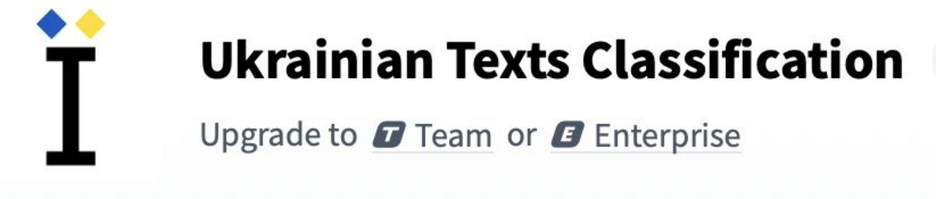
Using English as a proxy for a translation fielded worth results than using original Ukrainian data.

	Joy	Fear	Anger	Sadness	Surprise	None	Pr	Re	F1
Keywords UK	**0.30**	**0.15**	**0.08**	**0.21**	**0.15**	**0.25**	**0.27**	**0.25**	**0.26**
Keywords EN	0.17	0.05	0.01	0.18	0.08	0.11	0.15	0.01	0.10
UKR-RoBERTa-base UK	**0.65**	**0.58**	0.14	**0.50**	**0.49**	**0.74**	**0.56**	**0.49**	**0.52**
UKR-RoBERTa-base EN	0.53	0.24	**0.19**	0.30	0.31	0.60	0.32	0.42	0.36
mBERT UK	**0.46**	**0.24**	0.00	**0.45**	0.33	**0.73**	**0.38**	**0.38**	**0.37**
mBERT EN	0.38	0.12	**0.12**	0.31	**0.31**	0.55	0.31	0.30	0.30
LaBSe UK	**0.67**	**0.73**	**0.30**	**0.65**	**0.54**	**0.80**	**0.59**	**0.65**	**0.62**
LaBSE EN	0.60	0.41	0.22	0.39	0.30	0.64	0.44	0.43	0.43
XLM-RoBERTa Large UK	**0.73**	**0.79**	**0.20**	**0.68**	**0.60**	**0.80**	**0.61**	**0.68**	**0.63**
XLM-RoBERTa Large EN	0.50	0.34	0.15	0.47	0.24	0.53	0.33	0.45	0.37
Twitter-XLM-RoBERTa UK	**0.72**	**0.76**	0.13	**0.64**	**0.54**	**0.79**	**0.60**	**0.59**	**0.60**
Twitter-XLM-RoBERTa EN	0.62	0.26	**0.21**	0.52	0.44	0.62	0.42	0.47	0.44
Multilingual-E5 Large UK	**0.73**	**0.81**	**0.31**	**0.69**	**0.60**	**0.81**	**0.65**	**0.68**	**0.66**
Multilingual-E5 Large EN	0.61	0.26	0.22	0.36	0.23	0.56	0.36	0.41	0.37

Takeaways

Main takeaways

- **Multilingual encoders** worked better than Ukrainian-only monolingual ones;
- **LLMs prompting** surprisingly worked better then encoder-tuned models;
- Still, writing instruction in **prompts in English** gets better results than in other underrepresented target language.



Q&A

Questions?

Toloka | MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Case study

JEEM

VQA benchmark for four Arabic dialects

Kadaoui et al., Findings 2026 · slides by Karima Kadaoui

Case study

Motivation

VLMs today often struggle to generalize across **culturally diverse** and **dialect-rich** environments.

Current Vision Language resources

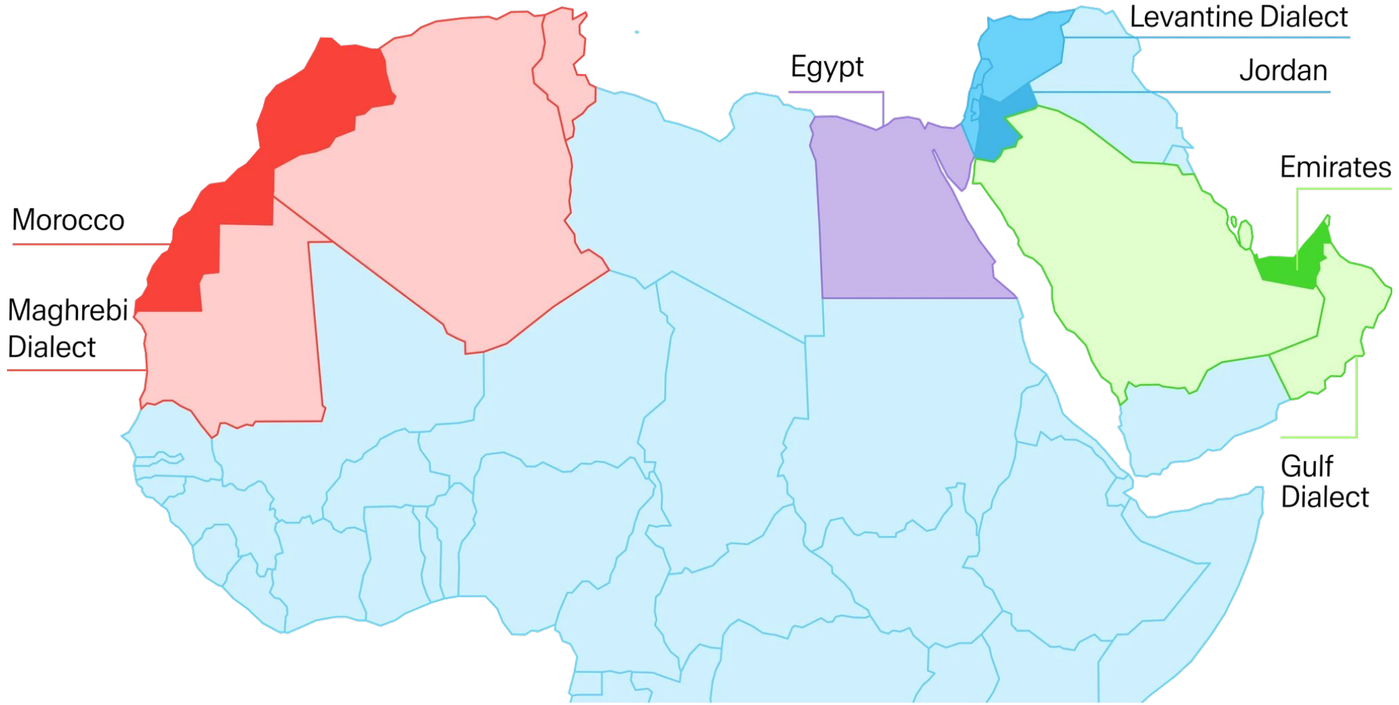
- Exclude Arabic
- Are western-centric benchmarks translated into Arabic
- Are MSA-only
- Include a single dialect
- Make use of synthetic questions
- Are text-only

New benchmark

Coverage

4 Dialects from
4 Countries/Regions:

🇯🇴 Jordanian	🇪🇬 Egyptian
🇦🇪 Emirati	🇲🇦 Moroccan



New benchmark

Two tasks

Image Captioning

MSA + Dialect

4,392 total

Visual Question Answering

5 visually-relevant QA pairs

10,890 pairs total

MSA Caption:

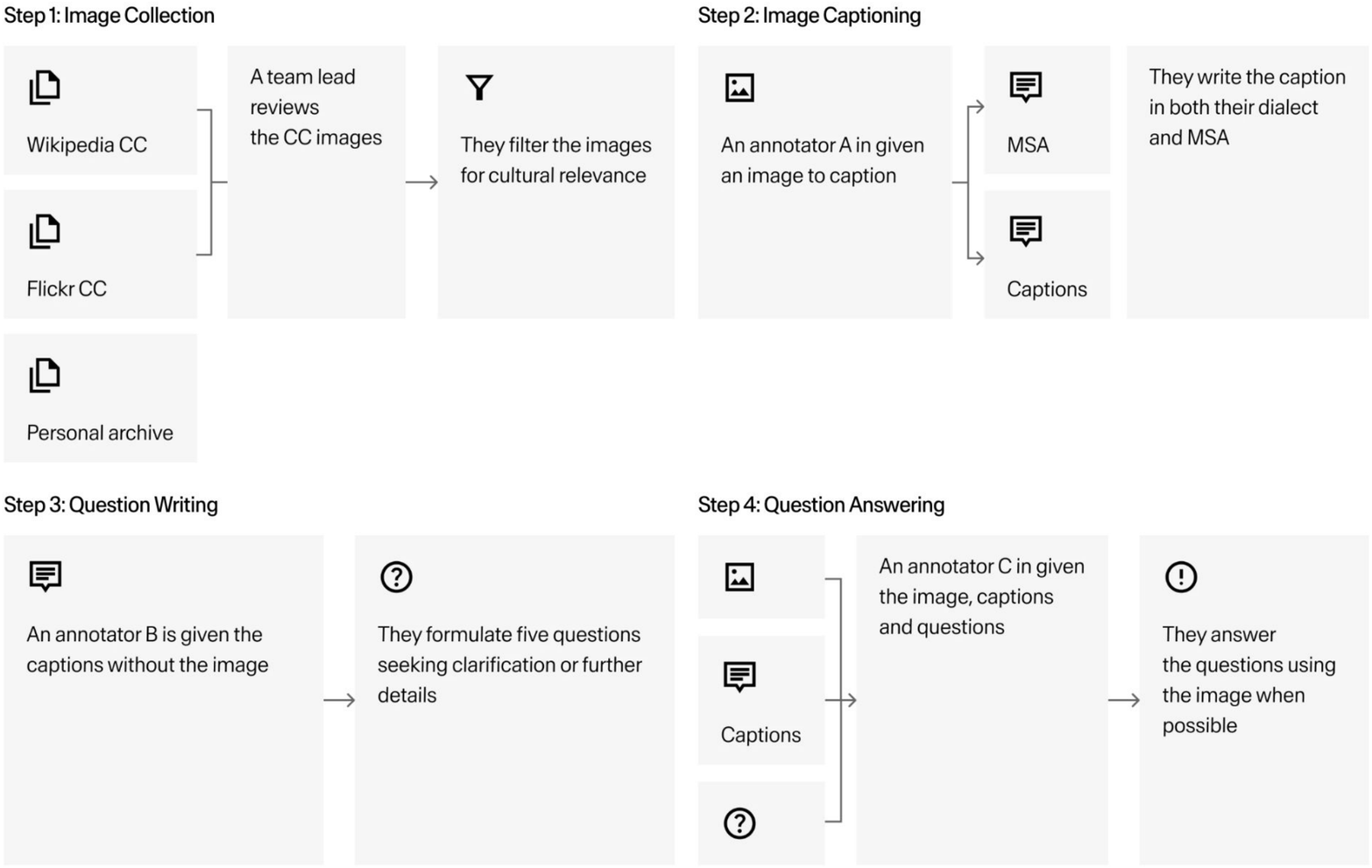
في الصورة يظهر رجل جالس على الأرض وهو يعزف على آلة موسيقية معروفة في المغرب باسم "الجمبري".
الرجل يرتدي عباءة حمراء مزخرفة، وعلى رأسه طربوش أحمر مزين كذلك. بجانب الرجل توجد حقيبة بنية قديمة بعض الشيء.

Dialectal Caption:

فالصورة كيبان واحد الراجل جالس فالأرض وكيعزف على آلة موسيقية لي معروفة فالمغرب بسمية "الجمبري"،
الراجل لابس واحد الغندورة حمرة مزوقة، وفراسو طربوش حمر حتى هو مزوق. حدا الراجل كاين واحد الصاك قهوي قديم شوية.

Pipeline

Data collection pipeline



Pipeline · Detail

Data collection pipeline

Similar projects can be implemented on the self-service Toloka platform.
Filter contributors by language proficiency · Pipeline of consecutive tasks, each step including creation and validation.

01

Image collection

Contributors collect culture-specific images from local contexts.
Independent reviewers verify that images are representative and do not contain sensitive content.

02

Caption writing

Contributors write captions in the target language.

Separate reviewers check captions for language quality, style, and relevance to the image.

03

Question writing

Contributors generate questions based on the captions.

Independent reviewers validate questions for language quality and stylistic consistency.

04

Answer writing

Contributors write answers grounded in the images

Separate reviewers verify answers for language quality and alignment with both the image and caption

05

Final validation

A final review step checks the entire data instance (image, caption, question, answer) for overall consistency, quality, and cultural appropriateness

Cultural aspects

Shared pool, diverging perceptions

- We include a shared pool of 100 images
- Demonstrates how cultural perspective shape perception
- E.g. The Omani halwa dessert only correctly identified by the Emirati annotators

Jordanian	Traditional dessert... almonds... pistachios... karawya or dibs	طبق حلو تقليدي... اللوز... الفستق... بالكراوية أو الدبس
Emirati	**Omani halwa**	حلوى عمانية
Egyptian	Pudding... chocolate... pine seeds	لبودنج... شيكولاتة... صنوبر
Moroccan	Chocolate... caramel... coconut and pistachios	شكلاط... كراميل... بالكوكو و بيسطاش

Benchmark

Benchmarking VLMs

6 models

Arabic-capable VLMs

AIN, AyaV, Maya, Palo, Peacock, GPT4o

4 metrics

Traditional

BLEU (B) · ROUGE (R) · CIDER (C) · BERTScore (BSc)

3 metrics

Additional

ALDi · DCScore · GPT-4-as-a-Judge

4 dimensions

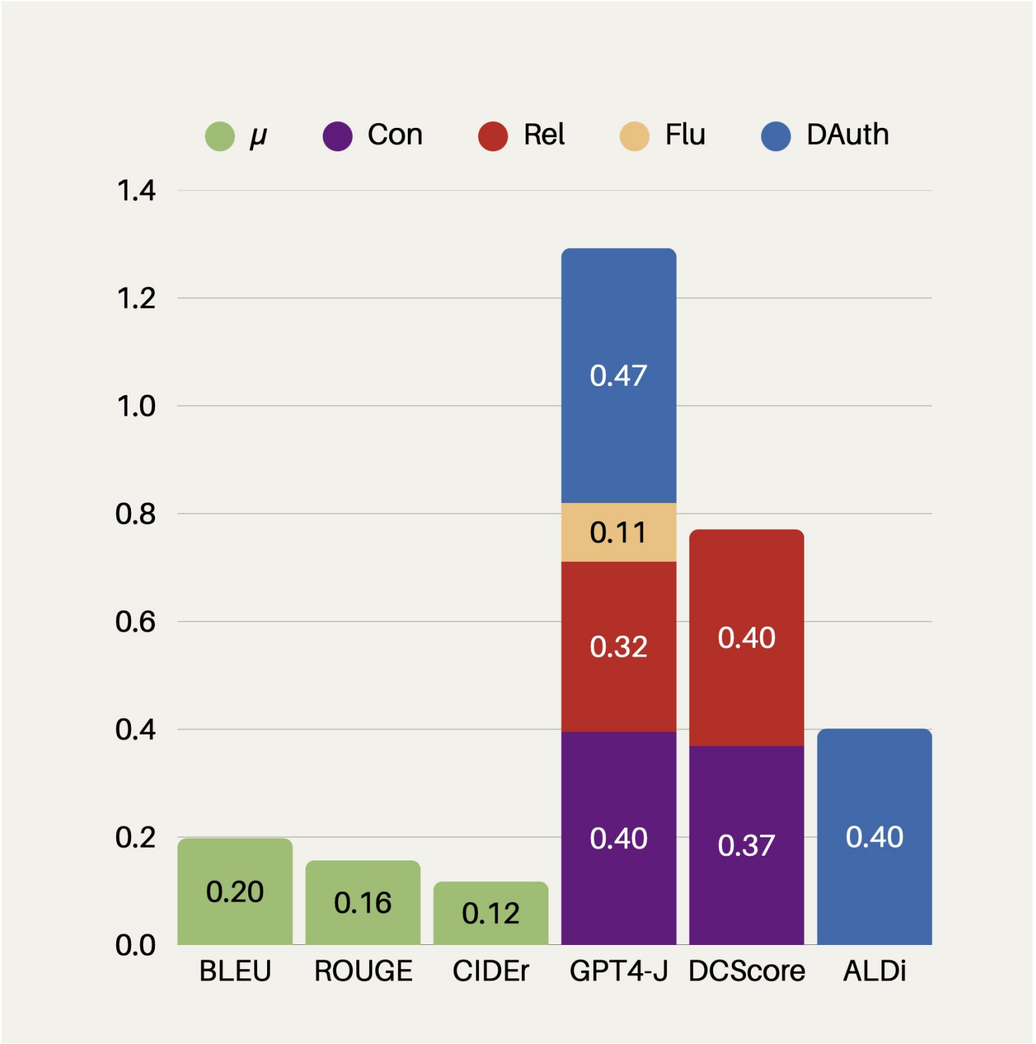
Evaluation

Consistency (Con) · Relevance (Rel) · Fluency (Flu) · Dialect Authenticity (DAuth)

Reliability

Metrics reliability

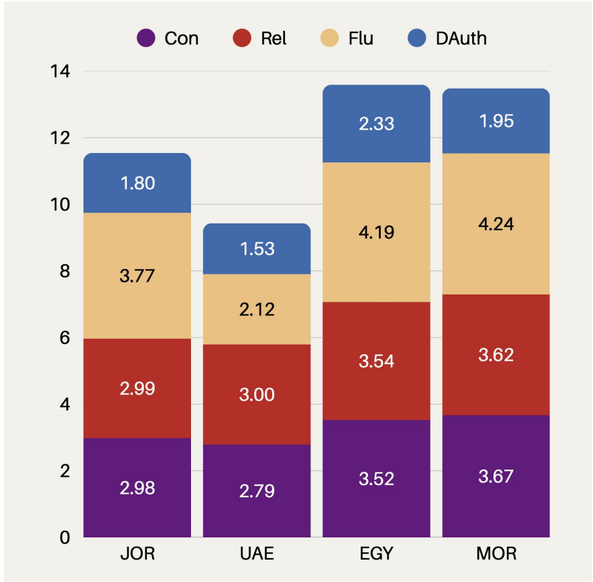
- Human eval of 350 images
  - 100 from each dialect (except UAE: 50)
  - 3 annotators/sample (except UAE: 1)
- Traditional metrics: weak alignment with human judgements.
- DCScore: highest correlation with Consistency and Relevance



Results

Image captioning

GPT-4o



Traditional metrics

- GPT-4o performs best across all metrics
- AyaV performs best open-source
- Rest of the models show quite similar scores
- BLEU-1 was used since higher-order BLEU scores were consistently near zero

Human evaluation

- Despite potential GPT-J bias, its ranking aligns with human judgement
- Humans place GPT-4o much higher at DAuth than ALDi
- 4 of the models struggle similarly with DAuth
- Humans judge Flu harsher than GPT-J.

GPT-as-a-Judge & DCSCore & ALDi

- Trends persist for GPT-4o and AyaV
- All models perform best at Fluency, and worst at Dialectal Authenticity
  - Likely due to generation in MSA
- DCScore shows slight differences to GPT-J
- ALDi places AyaV above GPT-4o

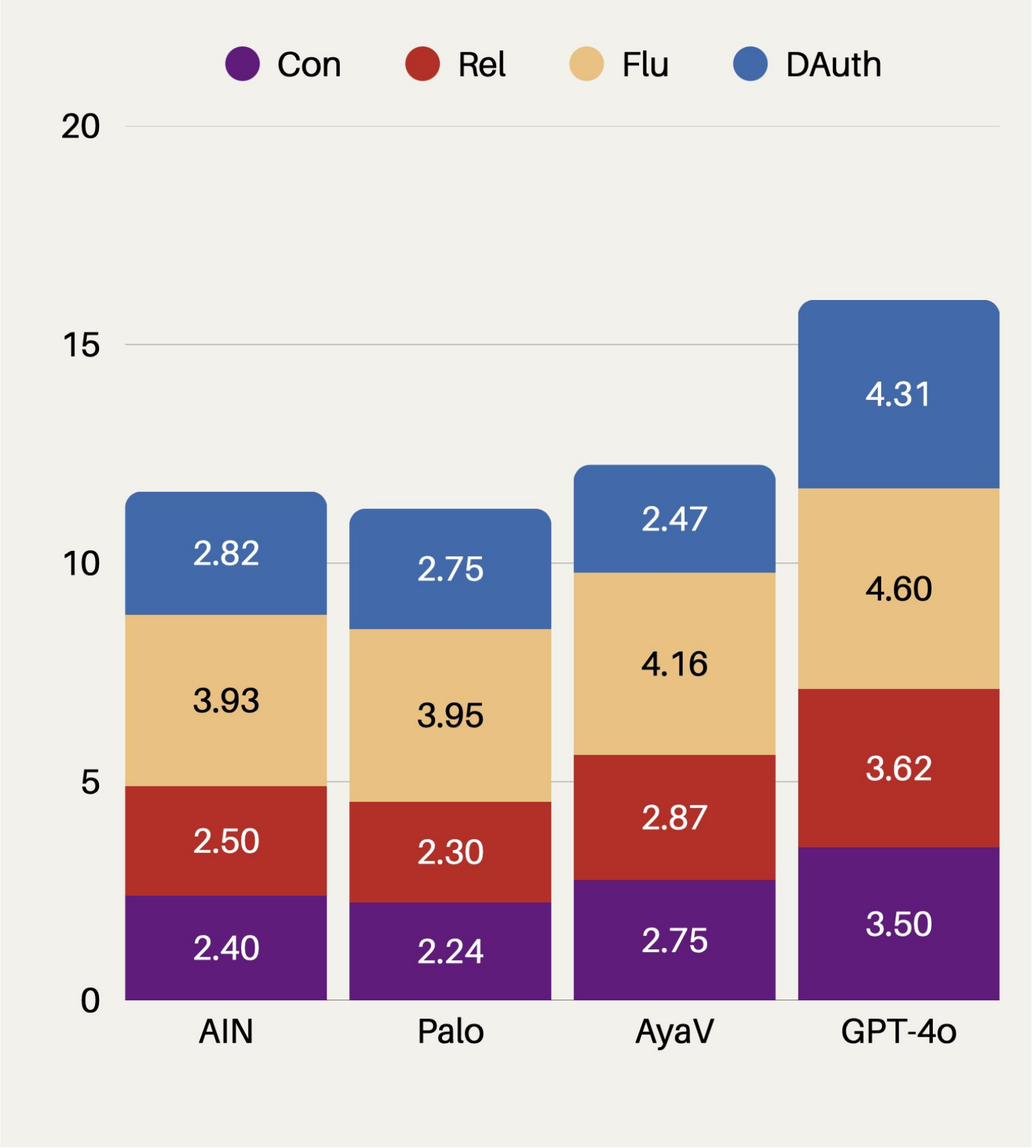
Dialect comparison

- Emirati is the trickiest across dimensions
- Models have better familiarity with Moroccan and Egyptian culture/dialects
- DAuth is the most challenging dimension

Results

Visual Question Answering

- Since multiple answers are possible, traditional metrics are not suitable
- GPT4-as-a-Judge evaluation shows similar trends to IC
  - Fluency is the highest-scoring dimension
  - Models struggle with Dialect Authenticity
  - Lower Consistency and Relevance scores indicate unsuccessful addressing of questions



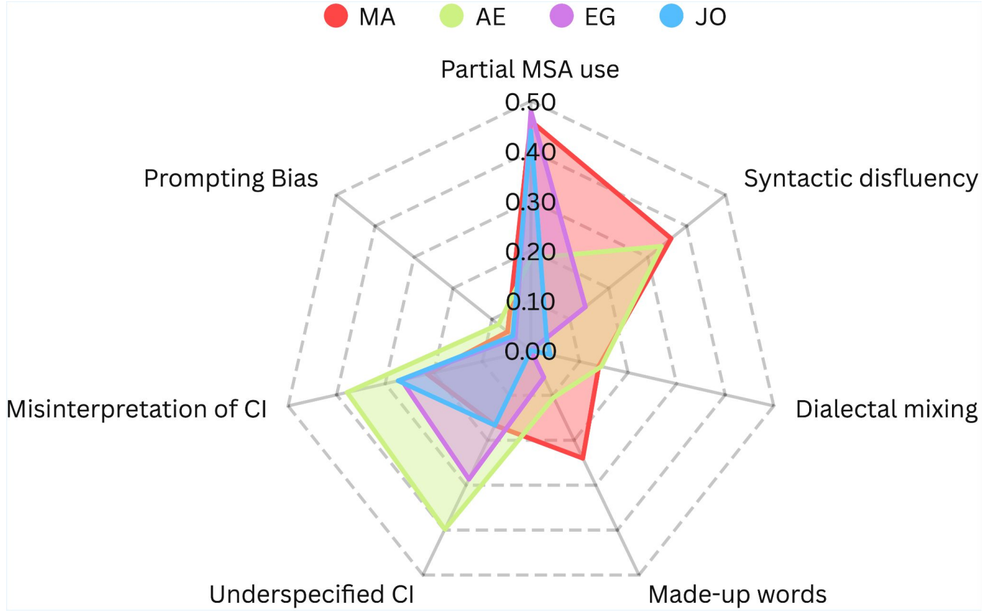
Analysis

Error analysis

(5 img from each country + 5 general Arab culture) × 6 models × 4 dialects = **600 evaluated predictions**

- Moroccan and Emirati sets are more challenging than Jordanian and Egyptian
- Culturally: Emirati most challenging
- Linguistically: Moroccan most challenging
- Similar degree of partial MSA use

Category	Type	Description
Dialectal	MSA use, partial	*Use of MSA words when dialectal alternatives exist.* مشهد رائع ديال غروب الشمس فوق المياه.
	MSA use, complete	*Use of MSA in the entire text.* الصورة تظهر مشهدًا في الشارع حيث يتم ركن سيارتين أمام مبنى.
	Made-up words	*Non-existent words that resemble dialect.* الأحمر، الأصفر، والأزرق كالت وحدات ملونة بألوان زاهية.
	Dialect mixing	*Multiple dialects mixed in one text.* الصورة دي كتعرض شجرة النخيل
	Syntactic dysfluency	*Awkward or incorrect sentence structure.* الجو حمابي بزاف، والسماء صافية، والجو كيفتح
Cultural	Underspecified CI	*Vague or imprecise mention of a cultural item.* آلة موسيقية instead of آلة الجمبري
	Misinterpretation of CI	*Misattributing a cultural item to a different culture.* تصوير يصور شخصًا يرتدي قبعة ملونة وملابس تقليدية مكسيكية
	Prompting bias	*Matches content to prompt's region, even if incorrect.*
Visual	Hallucination	*Mention of items not present in the image.* On an image with no birds: بالإضافة إلى ذلك، هناك عدة طيور
	Wrong count	*Incorrect number of items mentioned.* On an image of twelve men: رجلان يركبان على ظهور خيل سوداء
Generation Failure	Irrelevant info.	*Off-topic or unrelated content.* هاد الشجرة معروفة بفوائدها الصحية، وكتستخدم بكثرة للطعام والزيت
	Incomplete	*Ends generation abruptly.* في أحد شوارع المدينة المزدحمة
	Degeneration	*Repetitive wording or looping phrases.* هاد الصورة فيها جمل وكأنه كيف كيف كيف. وكأنه كيف كيف كيف...



Final

Main takeaways

- JEEM uses everyday images to encourage natural writing and create a realistic test bed
- JEEM offers a novel resource for evaluating and improving language models in real-world contexts.
- Even frontier models struggle with processing everyday languages, making them less accessible to many language communities.
- LLM-as-a-judge approaches correlate well with human judgments, offering scalable and reliable evaluation of VLM performance.

JEEM @ Hugging Face Datasets

Q&A

Questions?

Part 03

Expert Interviews

On Benchmark Creation in LR Settings

Ekaterina Artemova, Daryna Dementieva, **Shu Okabe**

Motivation

How do YOU collect data in LR languages?

Based on our own experience, we know that many themes remain behind the scenes.

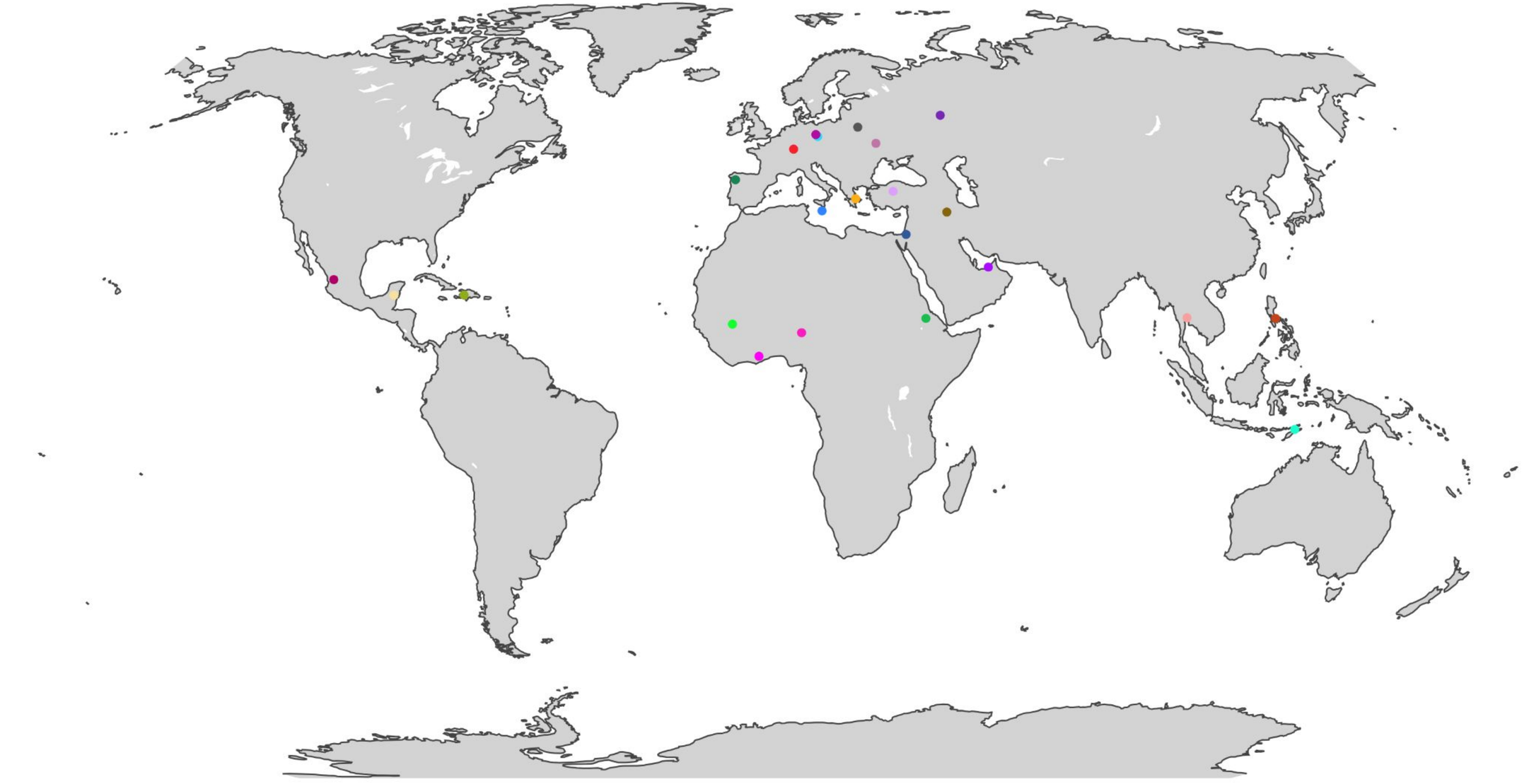
- What’s driving you, is it personal motivation or something else?
- How well do you really know the language you’re working with?
- Where do your annotators come from, and do you have the resources to support them?
- Do you stick to English-centric standards, or are you willing to rethink the rules?
- And how challenging is it, in practice, to get this work published?

Our approach

Study design

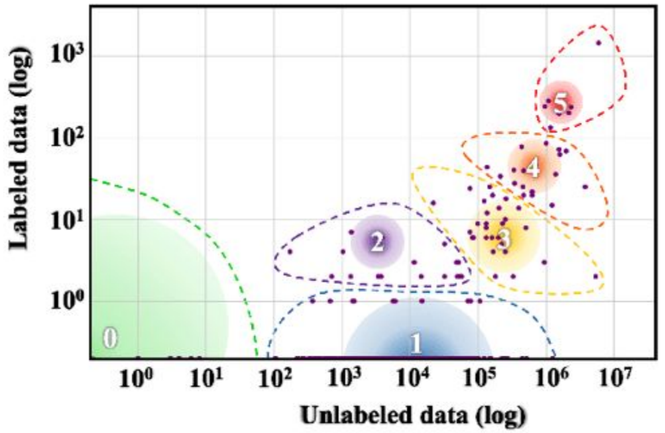
What really goes into building datasets across languages?

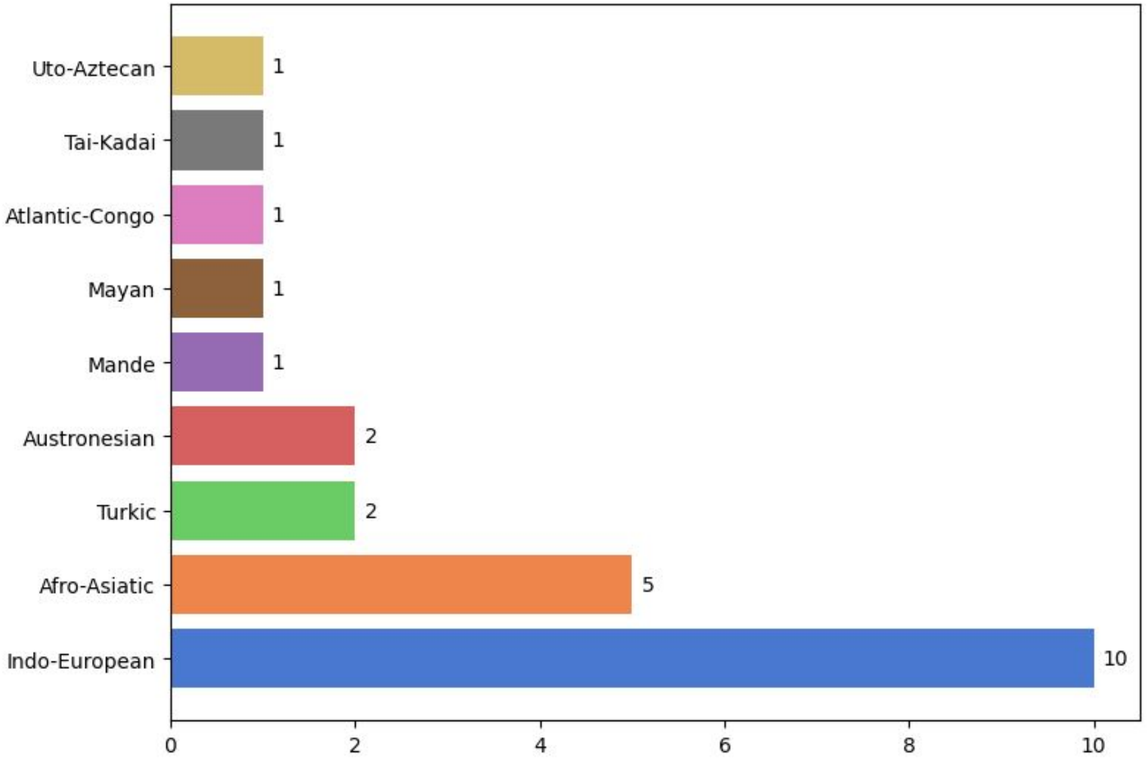
- We designed a structured questionnaire and conducted in-depth interviews with dataset authors
- 26 interviews were completed between September 2025 and April 2026
- The study covers languages from diverse cultures and geographic regions


Languages
Emirati (Gulf) Arabic
Maltese
Central Kurdish
Tatar
Belarusian
Galician
Malian languages Bambara
Yucatec Mayan
Irish
Tigrinya
Tagalog
Turkish
Twi
Hebrew
Greek
Thai
Haitian Creole
Alsatian
Tetun Dili
Upper Sorbian
Lower Sorbian
Ukrainian
Hausa
Wixárika

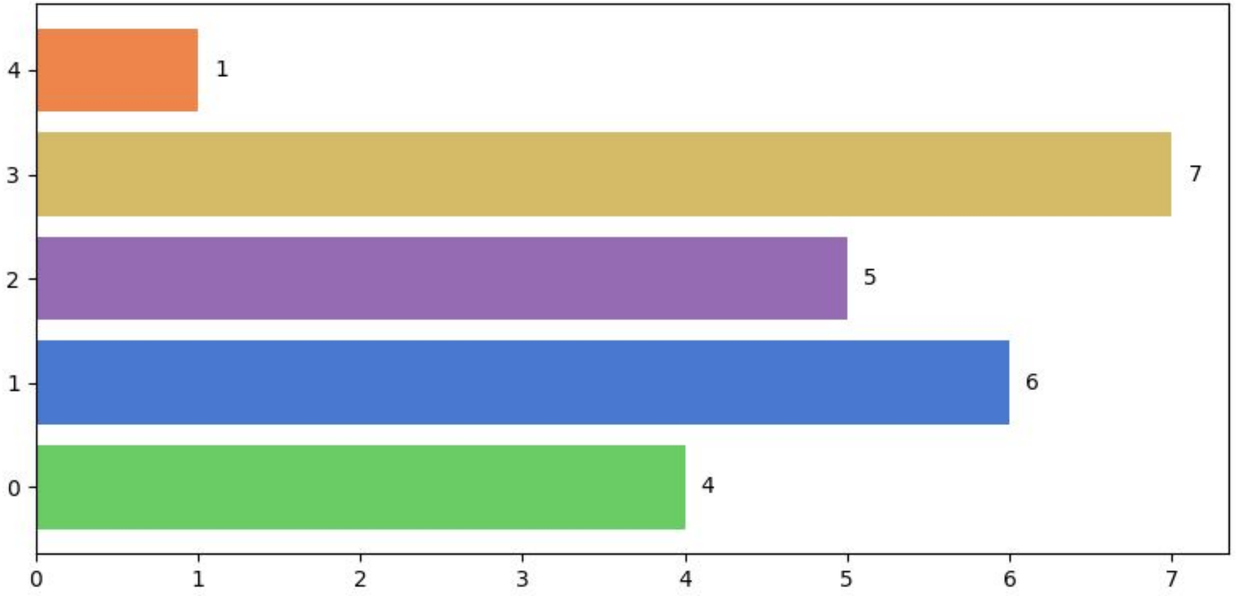
Our approach

Language coverage





NLP resource taxonomy from [Joshi et al. (2020)]

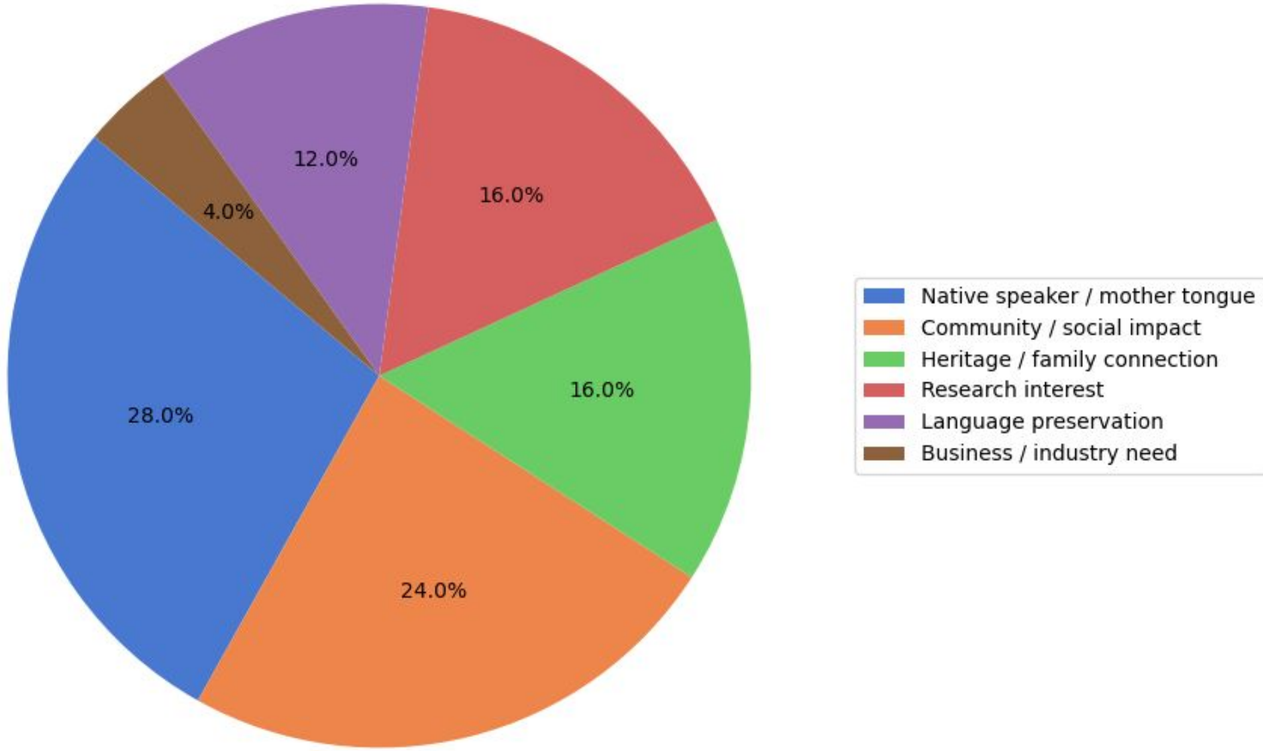


Findings

Motivation

What drives your interest in dataset collection?

- Personal connection
  - Native speaker (7)
  - Family connection (4)
- Sense of duty: social impact and preservation
  - I want my language to be represented (6)
  - I want to record my knowledge (3)
- Alignment with research topics (4)
- Connection to industry
  - There is a market for my language (1)

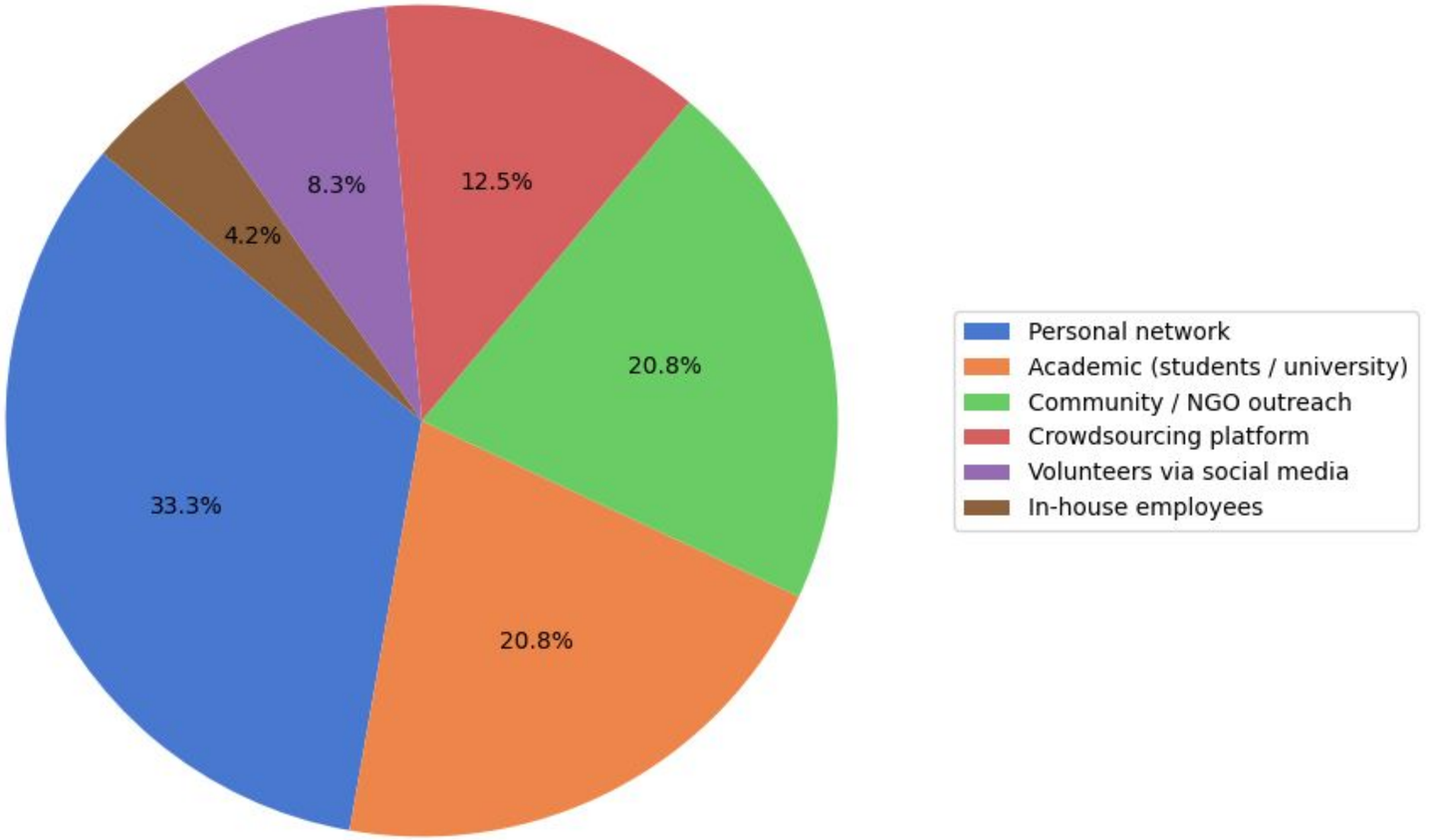


Findings

Annotation

Where do your annotators come from?

- Personal and academic network
  - Friends & family, past collaborators (8)
  - Local annotators from the university (5)
- Community outreach
  - Direct or indirect (5)
  - Social media (2)
- Hired experts
  - Crowd-sourcing (3)
  - In-house annotators (1)



Findings

Dataset design

Do you follow English-centric standards?

Partially yes (16)

- Translation of an existing benchmark
- Adaptation needed

No, built from scratch (8)

- No equivalent in English benchmarks
- Costly

Findings

Publishing

How difficult it is to get published?

No: welcoming (5)

- Better than earlier
- Dedicated submission track

Yes: somewhat (11)
& very difficult (8)

- High expectations
  - Difficult / impossible in practice

Call for action

We need to do better

- Continue **welcoming data collection initiatives** for low-resource languages
- Difficulty to get **support** and **funding** from institutions
- Multilingual benchmarks **tend to contain errors**, especially when there is no checks from native speakers
- Continue to **work with native speakers**

Q&A

Questions?

Final

Takeaways

Final

Takeaways

- **Four stages** in NLP pipeline:
  - **language identification** is still challenging
  - data **collection** and **annotation**
  - **parallel sentence mining** for MT
  - **combination** of techniques for downstream model acquisition: case studies
- **Working with native speakers is crucial**!
- **Progress towards data collection and annotation efforts** for low-resource languages

Thank you!

Acknowledgements

- **TUM** members have been funded by the European Research Council (ERC) under grant agreements No. 101113091 - Data4ML (a Proof of Concept Grant) and No. 101141712 - EPICAL.
  Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council.
  Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

European Research Council
Established by the European Commission