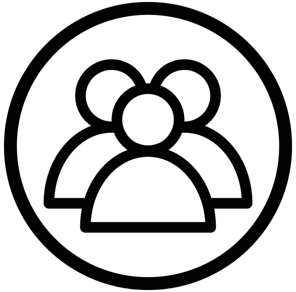
# SemRel 2024:

## A Collection of Semantic Textual Relatedness Datasets for 13 Languages

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

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



First team effort to build 13 sentence-based semantic textual relatedness datasets used in a SemEval shared task (>160 participants).

# Semantic Textual Relatedness (STR)

STR involves:

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

# Semantic Textual Relatedness (STR)

- STR is central to understanding meaning in text.

- Its applications include:
  - Evaluating sentence representation methods.
  - Question Answering.
  - Summarisation.
  - …

# Semantic Textual Relatedness (STR)

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

- Most people will agree that the sentences in pair 2 are **more related** than the sentences in pair 1.

# Semantic Textual Relatedness (STR)

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

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

# STR Data

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

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

# STR Data Creation

## Key Steps

1. Data selection → 2. Data annotation → 3. Quality control

**1. Data selection**

- Source data identification.
- Sentence Pairing

**2. Data annotation**

- Using comparative annotations (BWS: Best-Worst Scaling).

**3. Quality control**

- Dealing with Disagreements.
- Sanity check and postprocessing.

# STR Data Creation
## Data Selection

- Identify data sources (e.g., previously collected corpora, Wikipedia).
- Extract average-length sentences.
- Pair sentences to create instances.

# STR Data Creation

## Sentence Pairing

Random selection results in many unrelated sentences.

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

(High, medium, low, or unrelated).

# STR Data Creation

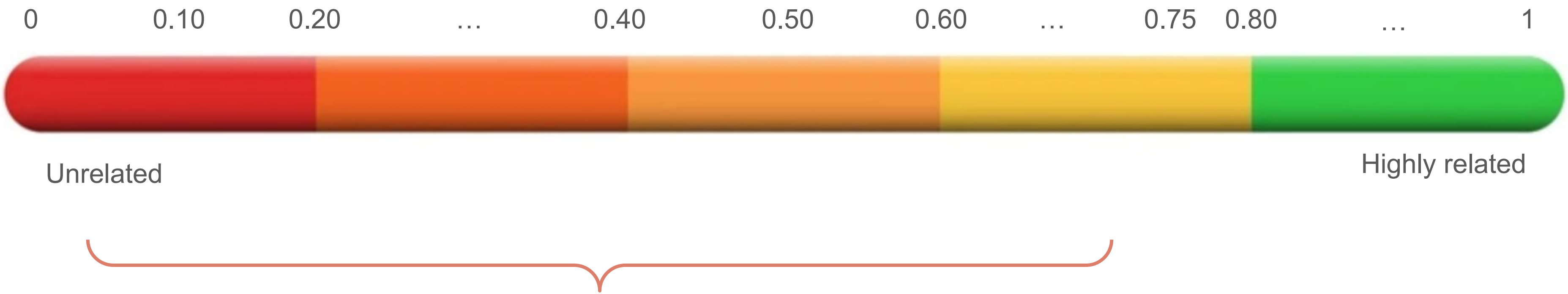
## Sentence Pairing

We build datasets within a wide range of relatedness scores.

# STR Data Creation
## Sentence Pairing Heuristics


**Lexical overlap** one or more words/tokens in common

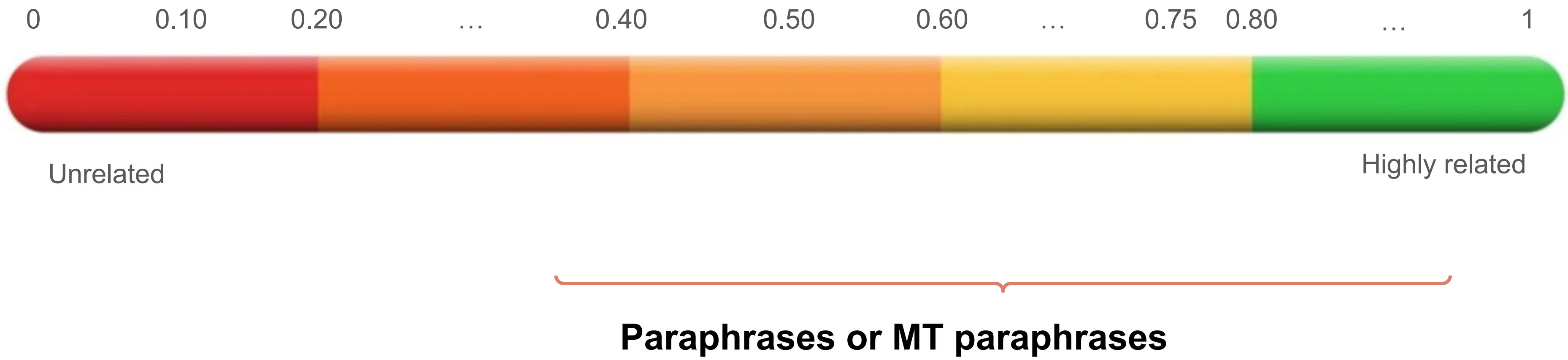
# STR Data Creation
## Sentence Pairing Heuristics



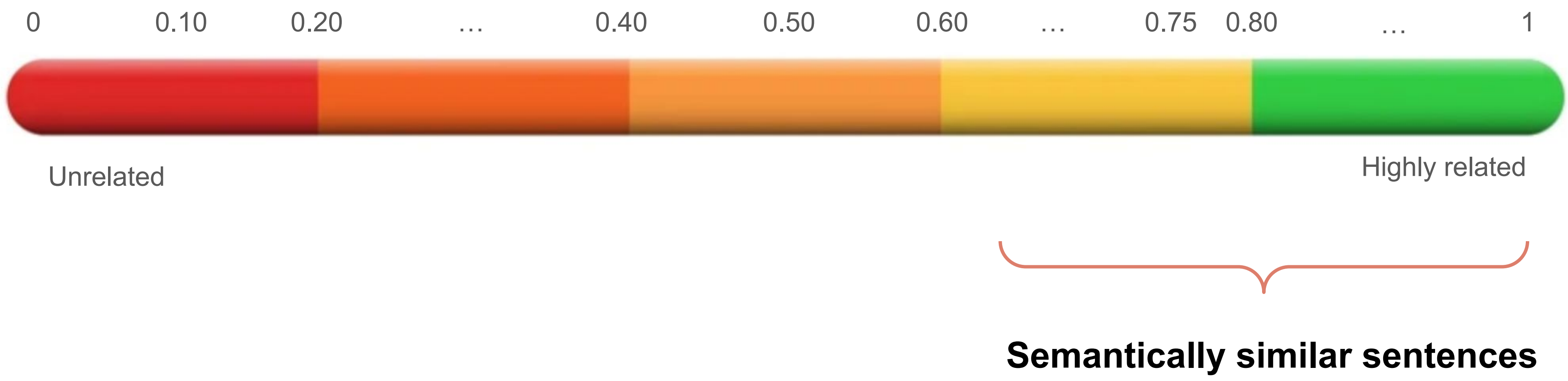

**Contiguity** sentences that appear one after the other

# STR Data Creation

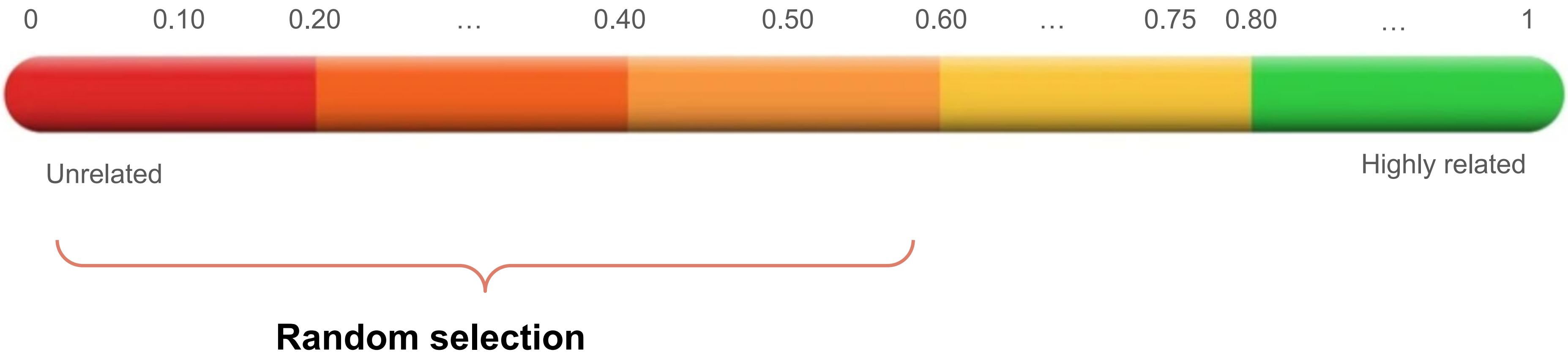
## Sentence Pairing Heuristics



# STR Data Creation
## Sentence Pairing Heuristics

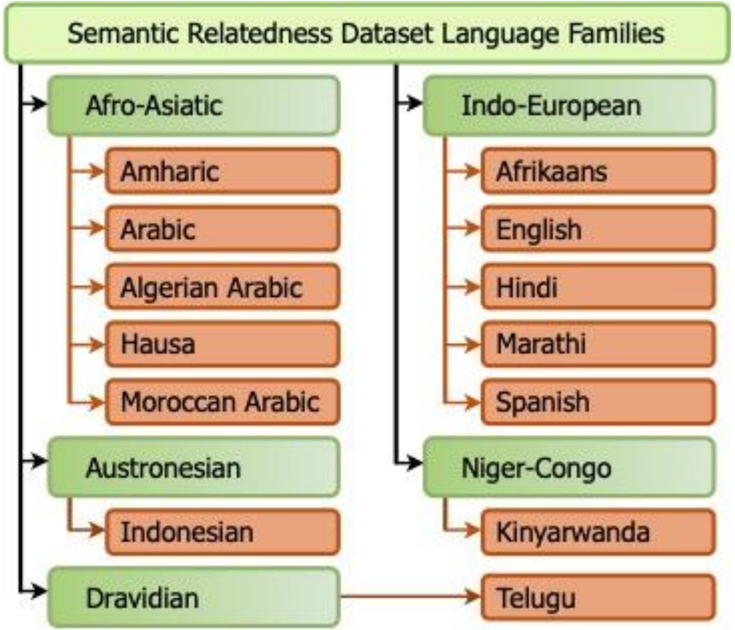


# STR Data Creation
## Sentence Pairing Heuristics



# STR Data Creation
## Languages

- 13 languages from 5 language families



# STR Data Creation

## Data Annotation

- We recruited native speakers

- We use comparative annotations (**BWS**: Best-Worst scaling):
    - **Compare** between pairs of sentences.
    - Choose the **Best** (most related) and the **Worst** (least related).

# STR Data Creation

## Data Annotation using BWS

Given **a tuple of 4 sentence pairs**: choose the **most related (best)** and the **least related (worst)** pair.

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

# STR Data Creation

## Data Annotation using BWS

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

- We rely on fluent speakers’ intuitions and avoid vague class definitions.
- We avoid biases of traditional rating scales.

# STR Data Creation

## Data Annotation using BWS

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

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

# STR Data Annotation

## Data Instances (1)

| L | Sentence #1 | Sentence #2 | Score |
|---|---|---|---|
| **Eng** | If that happens, just pull the plug. | If that ever happens, just pull the plug. | 1.0 |
| **Hau** | Haka ya furta a cikin jawabin sa na murnar cikar Najeriya shekaru 61 da samun ‘yanci. | Ya yi wannan iƙirarin e a cikin jawabin sa na murnar cikar Najeriya 61 da samun ‘yanci a ranar Juma’a. | 0.94 |
| **Amh** | መግለጫውን የተከታተለው የአዲስ አበባው ዘጋቢያችን ሰሎሞን ሙጬ ዝርዝር ዘገባ አለው ። | በስፍራው ተገኝቶ የተከታተለው የአዲስ አበባው ዘጋቢያችን ሰሎሞን ሙጬ ያጠናቀረውን ልኮልናል ። | 0.88 |
| **Ind** | Pendidikan Desa Pusaka memiliki 4 sekolah. | Pendidikan Desa Serumpun Buluh memiliki 4 sekolah. | 0.83 |
| **Arb** | في الواقع، هذه المادة التي ترون واضحة وشفافة. | مركبات هذه المادة هي فقط الماء والبروتين | 0.78 |
| **Ary** | وجدو راسكوم لرمضان.. الحرارة غادي تبدا بـ37 درجة فهاد المناطق | غير خرج رمضان وهي تشعل.. الحرارة غادي تبدا وغادي توصل لـ40 درجة فهاد المناطق | 0.75 |
| **Tel** | క్రికెట్ అన్ని ఫార్మాట్స్కు మలింగ గుడ్బై | కొలంబో: శ్రీలంక సీనియర్ పేసర్ లసిత్ మలింగ క్రికెట్ అన్ని రకాల ఫార్మాట్స్కు గుడ్బై చెప్పాడు. | 0.62 |

# STR Data Annotation

## Data Instances (2)

| L | Sentence #1 | Sentence #2 | Score |
|---|---|---|---|
| **Hin** | इस पर पीठ ने कहा कि इसे अब नौकरशाही पर नहीं छोड़ा जा सकता। | पीठ ने केंद्र की खिंचाई करते हुए कहा, आपके अधिकारियों ने कुछ नहीं किया है। | 0.5 |
| **Arq** | كاين واحد الأبيات يقولهم في الغنى تاعو تكوني تعرفيهم | اللي ما زهاش في الدنيا من الروح خالي | 0.5 |
| **Mar** | "ठाकरे सरकारच्या मंत्रिमंडळात 25 कॅबिनेट मंत्री असणार आहेत. | त्यामुळे गुढी पाडवा मेळाव्यामध्ये राज ठाकरे काय बोलणार याकडे सर्वांचे लक्ष लागून राहिले आहे. | 0.42 |
| **Esp** | ¿Qué país retiró sus tropas de Bosnia? | ¿Cuándo se ratificó la enmienda de sufragio femenino? | 0.23 |
| **Afr** | My eerste stukkie advies is dat jy realisties moet wees oor die afstand wat jy wil hengel. | Dit bring tot n einde die maanverkenningsprogram van die Verenigde State. | 0.19 |

# STR Data Creation

## Data Annotation using BWS: Reliability

## Split-Half Reliability Scores (SHR)

| L | afr | amh | arb | arq | ary | eng | esp | hau | hin | ind | kin | mar | tel |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ann/ tuple | 2 | 4 | 2-3 | 2 | 2 | 2-4 | 2-4 | 2-4 | 4 | 2 | 2 | 2-3 | 4 |
| train/ dev | 0.85 | 0.90 | 0.86 | 0.64 | 0.77 | 0.84 | 0.70 | 0.74 | 0.93 | 0.68 | 0.74 | 0.92 | 0.79 |
| Test | 0.85 | 0.90 | 0.86 | 0.64 | 0.77 | 0.80 | 0.70 | 0.74 | 0.94 | 0.68 | 0.74 | 0.96 | 0.96 |

# STR Data Creation

## Quality Control

- We inspected annotators with large **disagreements**
  - to ensure the annotation procedure was correctly followed.

# STR Final Datasets
## Disagreements

- We inspected annotators with large **disagreements**
    - to ensure the annotation procedure was correctly followed.

- **Sanity check**
    - Sentences with high relatedness scores had to be more semantically related than those with low relatedness scores.

# STR Final Datasets

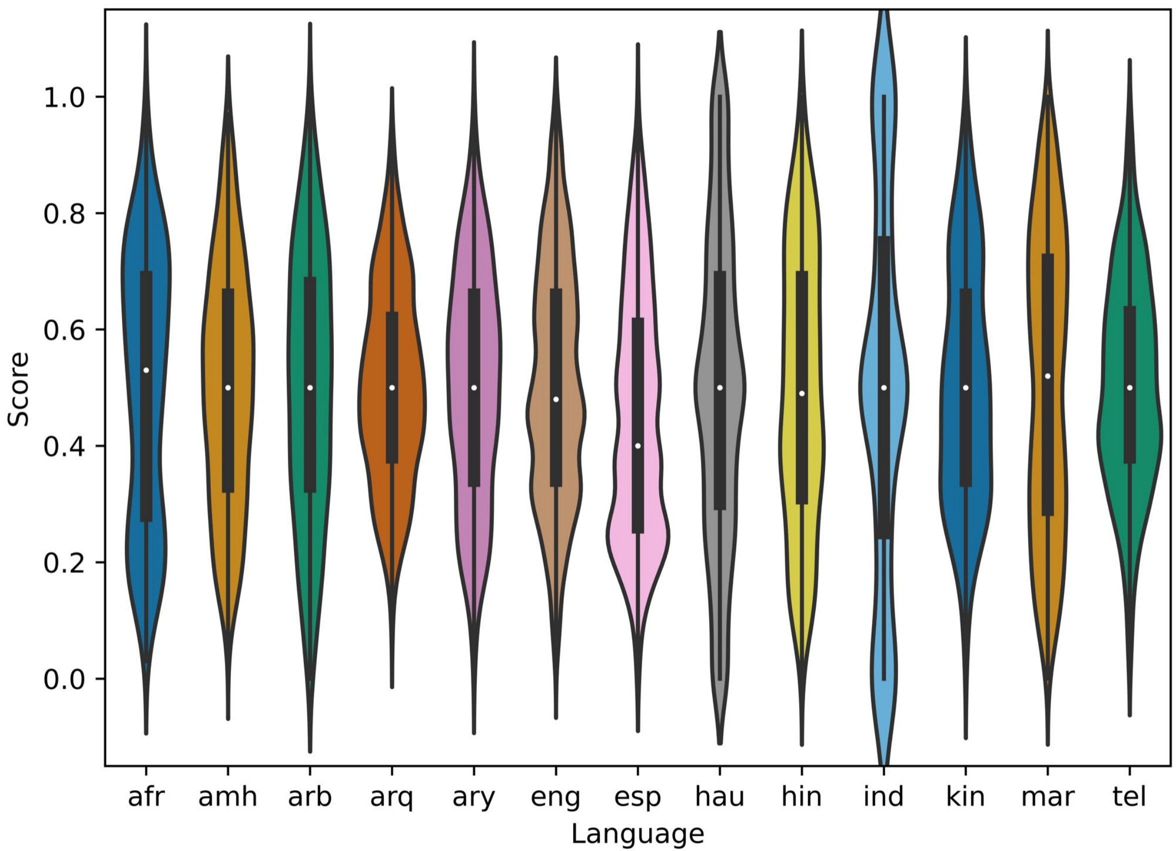
## Disagreements

- We inspected annotators with large **disagreements**
  - to ensure the annotation procedure was correctly followed.

- **Sanity check**
  - Sentences with high relatedness scores had to be more semantically related than those with low relatedness scores.

- **Postprocessing**
  - No repeated instances.
  - Text is well rendered and fully anonymised.
  - Control for expletives or inappropriate language.
  - Data is balanced.

# STR Final Datasets
## Distribution



# STR Final Datasets
## Data splits

| | afr | amh | arb | arq | ary | eng | esp | hau | hin | ind | kin | mar | tel |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Train** | - | 992 | - | 1,261 | 924 | 5,500 | 1,562 | 1,736 | - | - | 778 | 1,200 | 1,770 |
| **Dev** | 375 | 95 | 32 | 97 | 71 | 250 | 140 | 212 | 288 | 144 | 102 | 293 | 130 |
| **Test** | 375 | 171 | 595 | 583 | 426 | 2,600 | 600 | 603 | 968 | 360 | 222 | 298 | 297 |
| **Total** | 700 | 1,258 | 627 | 1,941 | 1,421 | 8,350 | 2,302 | 2,551 | 1,256 | 504 | 1,102 | 1,791 | 1,597 |

# Experiments

- Given sentence pairs, automatically determine relatedness scores.

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

# Experiments

- Given sentence pairs, automatically determine relatedness scores.

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

- **Metric** Spearman rank correlation coefficient.

# Experiments

## Settings

- **Supervised settings**
  - Train using the labeled training data.

# Experiments
## Settings

- **Supervised settings**
  - Train using the labeled training data.
- **Unsupervised settings**
  - Train without using any labeled STS or STR datasets between texts >2 words long in any language.

# Experiments
## Settings

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

# Experiments
## Settings

- **Supervised settings**
  - Train using the labeled training data.
- **Unsupervised settings**
  - Train without using any labeled STS or STR datasets between text >2 words long in any language.
- **Crosslingual settings**
  - Train without using any labeled STS or STR datasets in the target language.
  - Train using labeled datasets from 1 other language.
    - I.e., English for all non-English datasets and Spanish for the English dataset.
- **Note** Datasets without training sets (afr, arb, hin, ind) were only used in unsupervised and crosslingual settings.

# Experiments
## Models

- **Baseline**
    - **Lexical Overlap** number of unique unigrams occurring in sentences.

# Experiments
## Models

- **Baseline**
  - **Lexical Overlap** number of unique unigrams occurring in sentences.

- **Supervised**
  - **Multilingual** mBERT and XLMR for unsupervised settings.
  - **Monolingual** Language-specific LMs (e.g., BERTO, IndicBERT, DziriBERT, etc.).

# Experiments
## Models

- **Baseline**
    - **Lexical Overlap** number of unique unigrams occurring in sentences.

- **Supervised**
    - **Multilingual** mBERT and XLMR for unsupervised settings.
    - **Monolingual** Language-specific LMs (e.g., BERTO, IndicBERT, DziriBERT, etc.).

- **Unsupervised and Crosslingual**
    - LaBSE.

# Results

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

# Takeaways

Results show limitations of current models.

- E.g., mBERT for low-resource languages such as Hau and Amh.
- Language specific models did not always outperform multilingual ones.

Performance of current models are highly language-dependent

- It is not always related to high vs. low-resourcedness (e.g., MSA results).
- For low-resource languages, training data boosted the performance.

# Thank you!

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