SemRel2024: A Collection of Semantic Textual Relatedness Datasets for 13 Languages

Nedjma Ousidhoum[1]*, Shamsuddeen Hassan Muhammad[2]*, Mohamed Abdalla, Idris Abdulmumin[3], Ibrahim Said Ahmad[4], Sanchit Ahuja[5], Alham Fikri Aji[6], Vladimir Araujo[7], Abinew Ali Ayele[8,9], Pavan Baswani[10], Meriem Beloucif[11], Chris Biemann[8], Sofia Bourhim, Christine De Kock[12], Genet Shanko Dekebo[13], Oumaima Hourrane, Gopichand Kanumolu[10], Lokesh Madasu[10], Samuel Rutunda[14], Manish Shrivastava[10], Thamar Solorio[6], Nirmal Surange[10], Hailegnaw Getaneh Tilaye[15], Krishnapriya Vishnubhotla[16], Genta Winata[17], Seid Muhie Yimam[8], Saif M. Mohammad[18]

[1]Cardiff University, [2]Imperial College London, [3]Data Science for Social Impact Research Group, University of Pretoria, [4]Institute For Experiential AI, Northeastern University, [5]BITS Pilani, [6] MBZUAI, [7]KU Leuven, [8]Universität Hamburg, Language Technology Group, [9]Bahir Dar University, Faculty of Computing,[10]IIIT Hyderabad, [11]Uppsala University, [12]The University of Melbourne, [13]Adama Science and Technology University, [14]Digital Umuganda, [15]Kotebe University of Education, [16]University of Toronto, [17]HKUST, [18]National Research Council Canada

Contact: OusidhoumN@cardiff.ac.uk

Abstract

Exploring and quantifying semantic relatedness is central to representing language and holds significant implications across various NLP tasks. While earlier NLP research primarily focused on semantic similarity, often within the English language context, we instead investigate the broader phenomenon of semantic relatedness. In this paper, we present *SemRel*, a new semantic relatedness dataset collection annotated by native speakers across 13 languages: *Afrikaans, Algerian Arabic, Amharic, English, Hausa, Hindi, Indonesian, Kinyarwanda, Marathi, Moroccan Arabic, Modern Standard Arabic, Spanish,* and *Telugu*. These languages originate from five distinct language families and are predominantly spoken in Africa and Asia – regions characterised by a relatively limited availability of NLP resources. Each instance in the SemRel datasets is a sentence pair associated with a score that represents the degree of semantic textual relatedness between the two sentences. The scores are obtained using a comparative annotation framework. We describe the data collection and annotation processes, challenges when building the datasets, baseline experiments, and their impact and utility in NLP.

1 Introduction

Characterising the relationship between two units of text is an important component of constructing text representations. Within this context, semantic textual relatedness (STR) aims to capture the degree to which two linguistic units (e.g., words or sentences, etc.) are close in meaning (Mohammad, 2008; Mohammad and Hirst, 2012). Two units may be related in a variety of different ways (e.g., by expressing the same view, originating from the same time period, elaborating on each other, etc.). On the other hand, semantic textual similarity (STS) considers only a narrow view of the relationship that may exist between texts (such as equivalence or paraphrase) which does not incorporate other dimensions of relatedness such as entailment, topic or view similarity, or temporal relations (Abdalla et al., 2023; Agirre et al., 2013a). For example, *'I caught a cold.'* and *'I hope you feel better soon.'* would receive a low similarity score, despite the two being very related. In this work, we investigate the broader concept of semantic textual relatedness.

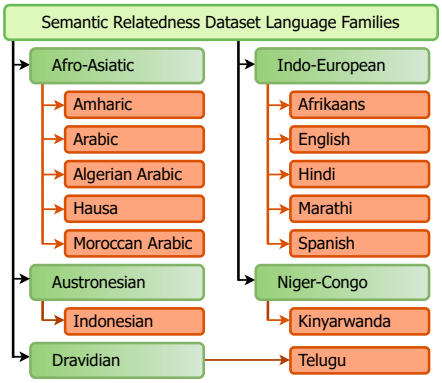

Figure 1: SemRel2024 languages and language families.

STR is central to understanding meaning in text (Hasan and Halliday, 1976; Miller and Charles, 1991; Morris and Hirst, 1991) and its automation can benefit various downstream tasks such as evaluating sentence representation methods, question answering, and summarisation (Abdalla et al., 2023; Wang et al., 2022).

Prior NLP work has mainly focused on textual

*Equal contribution from first and second authors, authors 3 to 26 are alphabetically ordered.

Findings of the Association for Computational Linguistics ACL 2024, pages 2512–2530

Lang.	Sentence 1	Sentence 2	Score
tel	ఇప్పటికే ధోనీ టెస్టు క్రికెట్ కు గుడ్ బై చెప్పిన విషయం తెలిసిందే.	అంతకు ముందు జరిగిన మరో రోడ్డు ప్రమాదంలో ఏడుగురు మరణించగా, 12 మంది గాయపడ్డారు.	0.02
afr	My eerste stukkie advies is dat jy realisties moet wees oor die afstand wat jy wil hengel.	Dit bring tot n einde die maan-verkenningsprogram van die Verenigde State..	0.19
esp	Costo monetario para mantener el microondas por $6.	Todavía nos quedan más de 200.000$ por recaudar de suscriptores y donantes como usted.	0.27
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
arq	كاين واحد الأبيات يقولهم في الغنى تاعو تكوني تعرفيهم	اللي ما زهاش في الدنيا من الروح خالي	0.50
arb	الآن انتقل بكم إلى لمحة عن تاريخ الاقتصاد و في نظري قد تكون مفيدة.	حوالي عام ١٨٥٠ كان صيد الحيتان من أكبر الصناعات في الولايات المتحدة	0.62
hin	देश में कोरोना वायरस से मौत का आंकड़ा 100 के पार पहुंचा, पिछले 12 घंटे में 26 की गई जान.	देश में कोरोना वायरस का कहर तेजी से बढ़ता जा रहा है।	0.72
kin	Duhugukire kwandika neza Ikinyarwanda Mu myandikire yIkinyarwanda hari amakosa akunda gukorwa ashingiye ku ifatana nitandukana ryamagambo.	Duhugukire kwandika neza Ikinyarwanda (igice cya gatatu) Mu myandikire yIkinyarwanda, hari amagambo afatana nandi atandukana.	0.75
ary	وجدو راسكوم لرمضان.. الحرارة غادي تبدا بـ٣٧ درجة فهاد المناطق	غير خرج رمضان وهي تشعل.. الحرارة غادي تبدا وغادي توصل لـ٤٠ درجة فهاد المناطق	0.75
ind	Pendidikan Desa Pusaka memiliki 4 sekolah.	Pendidikan Desa Serumpun Buluh memiliki 4 sekolah.	0.83
amh	**እኛን ከዚህ ጉዳይ ጋር የምንገናኝበት ቅንጣት ታክል ግንኙነት የለም**	**እኛን ቅንጣት ታህል ከዚህ ጉዳይ ጋ የሚያገናኘን ነገር የለም**	0.89
hau	Haka ya furta a cikin jawabin sa na murnar cikar Najeriya shekaru 61 da samun yanci.	Ya yi wannan iirarin e a cikin jawabin sa na murnar cikar Najeriya 61 da samun yanci a ranar Jumaa.	0.94
eng	I've been searching the entire abbey for you.	I'm looking for you all over the abbey.	1.00

Table 1: Examples of sentence pairs and their corresponding scores (from 0 to 1) in the various SemRel2024 languages. Examples are sorted by score and rows with higher degrees of relatedness are lighter colored. The translations can be found in the Appendix.

similarity, largely due to a dearth of relatedness datasets. Of the existing STR and STS datasets, most are in the English. The few STR and STS resources which exist for non-high resource languages are composed of word-level or phrase-level pairings. In this work, we curate 13 new monolingual STR datasets[1] for Afrikaans (`afr`), Amharic (`amh`), Modern Standard Arabic (`arb`), Algerian Arabic (`arq`), Moroccan Arabic (`ary`), English (`eng`), Spanish (`esp`), Hausa (`hau`), Hindi (`hin`), Indonesian (`ind`), Kinyarwanda (`kin`), Marathi (`mar`), and Telugu (`tel`).

The datasets are composed of sentence pairs, each assigned a relatedness score between 0 (completely unrelated) and 1 (maximally related). With the aim of curating diverse STR datasets, the pairs of sentences were first selected from pre-existing datasets covering various topics and formality levels, e.g., news data, Wikipedia, and conversational data. Additionally, we selected pairs with a large range of expected relatedness values by considering lexical overlap, contiguity, topic coverage, and random pairings. To generate the relatedness scores, the sentence pairs were then annotated by native speakers who performed comparisons between different pairs of sentences using Best-Worst Scaling (BWS) (Louviere and Woodworth, 1991). BWS is known to avoid common limitations of traditional rating scale annotation methods (Kiritchenko and Mohammad, 2016, 2017). The annotation process led to the high reliability of the final relatedness rankings in the different SemRel datasets. Our main contributions are as follows:

[1]Our team also created data for Punjabi. However, the sentence-pair selection procedure for it was markedly different than what we used for other languages. Therefore we do not include it here.

1. We present the first benchmark on semantic distance (similarity or relatedness) that includes low-resource African and Asian languages from five different language families (see Figure 1). Although Africa and Asia are home to over 5,000 languages from over 20 language families and have the highest linguistic diversity, there is little publicly available data on these languages.
2. We discuss general and language-specific challenges related to the data collection and annotation of the SemRel datasets.
3. We present baseline experiments conducted in different monolingual and crosslingual settings to demonstrate the usefulness and potential of our dataset collection.

To promote research in the field of semantic relatedness, we publicly released the SemRel2024 datasets as part of a shared task that attracted a large number of participants interested in low-resource languages.[2]

2 Related Work

The field of semantic relatedness in natural language processing covers a variety of approaches and techniques designed to measure the closeness in meaning between units, specifically words (Miller, 1994), or sentences (Abdalla et al., 2023).

Most prior work focuses on STS, a narrower subset of STR, and often only covers high-resource languages such as English (Agirre et al., 2012, 2013b, 2014, 2015, 2016; Marelli et al., 2014), Arabic, German, Spanish, Turkish (Ahmed et al., 2020; Cer et al., 2017b), and Italian (Glavaš et al., 2018) with the only exception being Finnish, Slovene, Croatian (Glavaš et al., 2018; Armendariz et al., 2020) and Farsi (Vulić et al., 2020). To overcome the scarcity of available resources, Tang et al. (2018) proposed sentence-level, encoder-based methods leveraging English data to create Arabic, Spanish, Thai, and Indonesian datasets, whereas Pandit et al. (2019) use traditional data augmentation methods to create Bangla data.

By comparison, this work is focused on the creation of resources for sentence-level STR in multiple low-resource languages. Here, the few works which exist for non-high-resource languages are at the word level (e.g., Yum et al., 2021 for Korean). To our knowledge, the only corpora specially designed for semantic textual relatedness between pairs of sentences was created by Abdalla et al. (2023) for English. Abdalla et al. (2023) curated a dataset of 5,500 English sentence pairs annotated using a comparative annotation framework. Their dataset has since been used to evaluate embedding approaches (Wang and Li, 2022) and other methods (Wang et al., 2022). The core of Abdalla et al. (2023) approach serves as the model for data annotations in this project. However, our work additionally explores new ways of data collection–curation, and several challenges had to be addressed when working with less-resourced languages.

3 STR Data

3.1 Data collection

A key step in the data creation process was identifying sources of text for each language and selecting sentence pairs. This was particularly challenging for low-resource languages such as Hausa, Kinyarwanda, and Algerian Arabic. Since arbitrarily selecting sentences and pairing them would lead to many unrelated instances, we relied on several heuristics, discussed in Section 3.1.1, to ensure a wide range of scores for each language. Since these methods are highly corpus- and language-specific, the approaches used per language were determined by native speakers. We provide the data origin and the pairing approaches used for each language in Section 3.1.2. The composition of the resulting dataset is summarised in Table 6 and the distribution of the relatedness scores across the datasets are illustrated in Figure 2.

3.1.1 Sentence pairing heuristics

Given a set of texts in a target language, careful consideration was given to the construction of sentence pairs to ensure that the pairs would exhibit relatedness scores varying from completely unrelated to very related. Since random selection would result in many unrelated pairs, we paired sentences mainly based on five methods previously defined by Abdalla et al. (2023) (described below). In cases where the pairs produced by the five methods were qualitatively judged to be insufficiently varied, we manually selected some instances to balance the data so that we have sufficient number of instances for each band of relatedness (high, medium, low, or unrelated).

[2] See https://semantic-textual-relatedness.github.io for more details.

Lang.	Curation technique	Data Sources
afr	Overlap, Random selection, Manual check	News data, reviews, recipes, blogs.
amh	Overlap, Similarity, Random selection, Manual check	News data, crawling.
arb	Overlap, Contiguity, Random selection, Manual check	Ted talk subtitles, news data.
arq	Overlap, Contiguity, Random selection, Manual check	YouTube comments, conversational data.
ary	Overlap	News data.
eng	Overlap, Similarity, Paraphrases, Contiguity, Randomness	Book reviews, news data, tweets, other.
esp	Overlap, Contiguity, Similarity	Movie reviews, news data, other.
hau	Overlap	News data.
hin	Overlap, Similarity, Contiguity, Paraphrase, Randomness	News data, other.
ind	Overlap	Wikipedia, news data.
kin	Overlap	News data.
mar	Overlap, Similarity, Contiguity, Paraphrase, Randomness	News data, other.
tel	Overlap, Similarity, Contiguity, Paraphrase, Randomness	News data, other.

Table 2: The curation techniques used for data creation. We list the main textual sources present in the datasets we used for instance creation. More details are shared in Section 3.1.2.

Lexical overlap Pairs are selected with various amounts of lexical overlap. That is, one or more words/tokens in common, with or without using TF-IDF normalisation. This method is expected to produce a wide range of relatedness values, and was used in most low-resource languages.

Contiguity/Entailment We select pairs of sentences that appear one after the other in a paragraph or a social media thread. This method is likely to produce pairs of sentences that are somewhat related and can contribute to representing the low to medium ranges of relatedness.

Paraphrases or Machine Translation (MT) paraphrases This method consists of selecting pairs of sentences from paraphrase or MT data. For MT, we pivot across the translation and back to the source language to generate a new sentence and pair it with the original. However, many low-resource Asian and African languages lack reliable MT resources.

Semantically similar instances Semantically similar sentences are selected from a publicly available dataset such as the STS dataset by Cer et al. (2017a) or manually identified by a native speaker in order to include highly related instances and balance the dataset.

Random selection Random sentences are selected. This method is expected to represent the low to medium ranges of relatedness.

Manual check In cases where the pairs produced by the above methods were qualitatively judged to be insufficiently varied, instances were manually selected to balance the data so that there were sufficient number of instances for each band of relatedness (high, medium, low, or unrelated). This can apply to any range of relatedness (i.e., high, medium, low, or unrelated).

3.1.2 Data curation

Since most of the SemRel languages are low-resource, the domain, (in)formality, and diversity of the sentence pairs were highly dependent on the publicly available corpora. We aimed to collect datasets with average-length sentences, free of offensive utterances, and as diverse as possible. As such, data instances were extracted for each language using a tailored combination of the heuristics described in Section 3.1.1. We used further pre-processing, post-processing, and data analysis methods (discussed below) to avoid incoherence and unnaturalness.

English and Spanish As English and Spanish are high-resource languages, we sampled sentences from various sources to capture a wide variety of sentence structure, formality, and grammaticality in texts. As shown in Table 6, we paired sentences in a number of ways that include lexical overlap, entailment, similarity and paraphrases. The English dataset includes sentences that have the same meaning but a different formality collected from the Formality dataset (Rao and Tetreault, 2018), tweets (Mohammad et al., 2017), paraphrases from machine translation systems extracted from the ParaNMT dataset (Wieting and Gimpel, 2017), book reviews from Goodreads (Wan and McAuley, 2018), pairs of premises and hypotheses from the SNLI dataset (Bowman et al., 2015), and semantically similar sentences (Cer et al., 2017a).

Similarly, we select pairs of Spanish sentences

Language	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
SHR 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
SHR 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

Table 3: SHR (split-half reliability) scores for each of the dataset splits and numbers of unique annotations per tuple (#Ann/tuple). As some languages (eng, hin, mar, and tel) had splits annotated in separate annotation efforts (instead of one combined annotation), we report the SHR scores for both.

from semantic similarity datasets such as STS (Agirre et al., 2014, 2015; Cer et al., 2017a), entailment datasets such as SICK-es (Huertas-Tato et al., 2021) and NLI-es (Araujo et al., 2022), and paraphrasing datasets such as PAWS-X (Yang et al., 2019). We also sampled contiguous sentences from XL-Sum (Hasan et al., 2021) and BSO DiscoEval Spanish (Araujo et al., 2022), and we included questions of different types from Spanish QC (Á. García Cumbreras et al., 2006).

Arabic Variations: Modern Standard Arabic, Algerian, and Moroccan Arabic Arabic is known for diglossia (Ferguson, 1959), meaning that Arabic varieties are used for different contexts. For instance, Modern Standard Arabic is usually used in formal and academic communication while dialects are typical for conversational settings. The various sources of the Arabic data are somewhat reflective of the distinct language usage scenarios.

Therefore, for Modern Standard Arabic (MSA), we used two datasets from different domains: TED Talk subtitles (Zong, 2015) on science, society, and art and news articles on economics (Al-Dulaimi, 2022). In addition to sentences with lexical overlap, we selected contiguous sentences in Ted Talk subtitles to include different degrees of relatedness, and as some sentences in the subtitles were slightly ungrammatical, we corrected them based on the standard Arabic grammar rules. For Algerian Arabic, we used CalYou (Abidi et al., 2017), a dataset composed of YouTube comments collected from major Algerian YouTube channels by 2017, and the Algerian instances spoken in two major Algerian towns (Algiers and Annaba) present in PADIC: Parallel Arabic Dialect Corpus (Meftouh et al., 2015). We used lexical overlap to pair sentences, picked contiguous ones in a conversation in PADIC, and added randomly or manually selected sentence pairs to balance the relatedness score distribution in the dataset. For both MSA and Algerian Arabic, we allowed short sentences as Arabic is highly inflectional. For Moroccan Arabic, we used headlines from the Goud.ma dataset introduced by Issam and Mrini (2022) and the Moroccan Arabic sentences were paired based on lexical overlap.

Afrikaans The Oscar dataset (Ortiz Suárez et al., 2020) was used as basis for the Afrikaans corpus. We chose sentences from news articles, blogs, reviews, and recipes. We also excluded sentences from religious texts and academic articles after observing that these did not produce high-quality pairs. We further excluded a number of advertorial texts that appear to be low-quality translations. All instances were then manually assessed for grammar and ungrammatical sentences were discarded. Sentences were paired if they had an overlap of at least five tokens and at least three non-overlapping tokens with matches within the same article only. Random sentence pairs were also included to calibrate the dataset.

Amharic, Hausa, Kinyarwanda For Amharic, we paired sentences present in news articles from different Ethiopian news outlets (Yimam et al., 2021). Similarly, the Hausa and Kinyarwanda datasets include pairs of sentences from news articles collected by Abdulmumin and Galadanci (2019) and Niyongabo et al. (2020), respectively. Sentences shorter than five words and longer than 20 were excluded, and pairs were created using lexical overlap. Additionally, for Amharic, we excluded sentences with mixed languages to avoid confusing the annotators.

Indonesian For Indonesian, we collected sentences from Wikipedia texts present in the ROOTS split (Laurençon et al., 2022; Setya and Mahendra, 2018) and the IndoSum (Kurniawan and Louvan, 2018) datasets. IndoSum is a human-written summarization dataset consisting of pairs of news articles with abstractive summaries. We parsed both corpora at a sentence level and only selected sentences that were composed of five to fifteen words.

Hindi, Marathi, and Telugu As these languages lack publicly available resources, especially Marathi and Telugu, we used the Mukhyansh

dataset only (Madasu et al., 2023) to create the sentence pairs. It is composed of news headlines and their corresponding articles and is diverse in nature. We created instances using lexical overlap, paraphrase generation, contiguous sentence selection, and random sentence selection to balance the data.

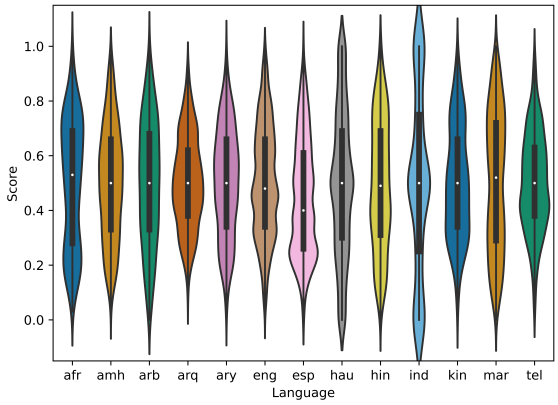

Figure 2: Violin plots representing the distributions of the relatedness scores. For instance, the distribution of the Arabic (`arb`) dataset is unimodal, Marathi's (`mar`) is bimodal, and the Indonesian (`ind`) dataset's is trimodal.

3.2 Data annotation and challenges

Annotation process Similarly to Abdalla et al. (2023), we used BWS to annotate our data instances and generate an ordinal ranking of instances[3]. Although pairwise comparisons are more reliable than simply labelling the sentence pairs as related or unrelated, it is a time-consuming process if performed on a large dataset as it requires $N \times N = N^2$ comparisons if performed on a dataset of N instances. Best-worst scaling mitigates this issue according to Kiritchenko and Mohammad (2017) as it leads to reliable scores from about $2 \times N$ comparisons of 4-instance tuples.

BWS requires fewer labels (Louviere and Woodworth, 1991), in our case, given four instances (i.e., pairs of sentences) p_i with $0 \leq i < 4$, for a tuple: $\langle p_0, p_1, p_2, p_3 \rangle$, if p_0 is marked as most related and p_3 as least related, then we know that $p_0 > p_1$, $p_0 > p_2$, $p_0 > p_3$, $p_1 > p_3$, and $p_2 > p_3$ ($<$ and $>$ refer to less and more related, respectively). We then use these inequalities to compute real-valued scores that consist of the fraction of times a pair p_i was chosen as the most related minus the fraction of times p_i was chosen as the least related. Then, an ordinal ranking of sentence pairs is generated (Orme, 2009; Flynn and Marley, 2014).

[3]The tuples were generated using the BWS scripts provided by (Kiritchenko and Mohammad, 2017): `http://saifmohammad.com/WebPages/BestWorst.html`.

Furthermore, the notions of *related* and *unrelated* have fuzzy boundaries with no singular accepted definition in the literature. Different people and different language cultures may have several intuitions of where such a boundary exists. Therefore, by using comparative annotations and relying on the intuitions of fluent speakers for each language to choose between sentence pairs, we can avoid ill-defined categories. This is in line with our goal of capturing common perceptions of semantic relatedness (i.e., what is believed by the vast majority) instead of "correct" or "right" rankings.

Instructions We selected native speakers to annotate the sentence pairs. Then, given a set of four sentence pairs, annotators were tasked with reporting on their relative relatedness. Concretely, given 4 sentence pairs, each of the form *[sentence A, sentence B]*, the task was to select the sentence pair that is the *most related* (i.e., sentence A is closest in meaning to sentence B) and the sentence pair that is the *least related* (i.e., sentence A is farthest in meaning to sentence B). The full instructions can be found in the Appendix.

In the guidelines, it was noted that sentence pairs that are more specific in what they share tend to be more related than sentence pairs that are only loosely about the same topic. Furthermore, if one or both sentences have more than one interpretation, the annotators have to consider the closest meanings.

Overall, by manually examining the annotations, we noted that the BWS framework does lead to more robust annotations. However, a downside is the fact that annotating one instance could take more than one minute and the task can be challenging since many instances to be compared can be similarly (un)related.

Annotation reliability In Table 3, we report the number of annotators and the split-half reliability (SHR) (Cronbach, 1951; Kuder and Richardson, 1937) scores for each of the datasets. SHR measures the degree to which repeating the annotations results in similar relative rankings of the instances. First, it splits the 4-tuple annotations into two bins. Then, the annotations for each bin are used to generate two different independent relatedness scores, and the Spearman correlation between the two sets of scores is calculated to estimate the closeness of the two rankings. A high correlation indicates that the annotations are reliable. This process is

Data	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	1,170
Test	375	171	595	583	426	2,600	600	603	968	360	222	298	297
Dev	375	95	32	97	71	250	140	212	288	144	102	293	130
Total	700	1,258	627	1,941	1,421	8,350	2,302	2,551	1,256	504	1,102	1,791	1,597

Table 4: Number of instances in the training, dev, and test sets for the different datasets. The languages with no training data (afr, arb, hin, ind) were only used in unsupervised and cross-lingual settings.

repeated 1,000 times and the correlation scores are averaged similarly to Abdalla et al. (2023). Overall the scores in Table 3 vary between 0.64 and 0.96, which indicates a high annotation reliability.

Disagreements We inspected annotators with large disagreements to ensure the annotation procedure was correctly followed (i.e., their annotations made sense for native speakers). Very strong disagreements would serve as a red flag of poor data quality resulting in a more thorough review of the annotation quality. Hence, as a sanity check, we examined whether sentences with high relatedness scores were more semantically related than those with low relatedness scores. The specific procedure for ensuring data quality depends on the annotation procedure of the team (e.g., those using AMT vs those who did not). Note that disagreements were not deleted, as they can serve as a useful signal for BWS. That is, for a tuple $\langle p_0, p_1, p_2, p_3 \rangle$, when annotators disagree on what is most related (e.g., p_1 or p_3), then it is an indication that p_1 and p_3 may be semantically close to each other. As all tuple annotations (twice the number of instances) are used to determine the final scores of the sentence pairs, this disagreement would lead to the two pairs (p_1 and p_3) getting scores that are close to each other. On the other hand, if a sentence pair consistently occurs in 4-tuples that have very low annotator agreement, then it is likely that the sentence pair is the source of disagreement. This can be due to various reasons such as the language use, code-switching, or the annotator's familiarity with the topics discussed.

Besides sharing our datasets with the community, we also make the full 4-tuple annotations public.

3.3 Postprocessing and data quality control

Quality control For our final dataset, we carried a data post-processing step to ensure that:

- no instances are repeated;
- the data does not include invisible characters, incorrectly rendered emoticons, or garbled encoding characters;
- texts are fully anonymised (deleting emails and IDs if they occur, replacing @mentions with @<username>, and replacing any URLs with non-identifiable placeholders);
- the data does not include a high amount of expletives or inappropriate language; and
- the data is balanced.

Manual Spotchecks Finally, a team of native speakers manually spot-checked the scores to make sure that the relatedness scores made sense and to supplement the quantitative evaluation based on SHR.

4 Experiments

4.1 Data

We use the splits reported in Table 4. For the languages without training data (afr, arb, hin, ind), we only report experiments in unsupervised and crosslingual settings. For English, we use the STR-2022 dataset (Abdalla et al., 2023) for training and we use our newly created dataset for testing.

4.2 Experimental setup

We report the Spearman correlation scores between the predicted labels and the gold standard ones for the different languages in three main settings:

- **Supervised** systems trained on the labeled training datasets provided.
- **Unsupervised** systems developed without the use of labeled datasets pertaining to semantic relatedness or semantic similarity between units of text of more than two words long in any language.
- **Crosslingual** systems developed without the use of any labeled semantic similarity or semantic relatedness datasets in the target language and with the use of data from at least one other language included in SemRel.

In our experiments, we use:

- a simple baseline based on the number of shared words (lexical overlap),

	afr	amh	arb	arq	ary	eng	esp	hau	hin	ind	kin	mar	tel
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 (Multilingual)												
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
	Unsupervised (Monolingual)												
AfroXLMR	0.45	0.40	0.18	-	-	0.30	-	0.07	-	-	0.16	-	-
ALBETO	-	-	-	-	-	-	0.62	-	-	-	-	-	-
AmRoBERTa	-	0.72	-	-	-	-	-	-	-	-	-	-	-
ARBERT	-	-	0.56	-	-	-	-	-	-	-	-	-	-
arb BERT	-	-	0.31	-	-	-	-	-	-	-	-	-	-
BETO	-	-	-	-	-	-	0.68	-	-	-	-	-	-
DziriBERT	-	-	-	0.43	-	-	-	-	-	-	-	-	
Indic-BERT	-	-	-	-	-	-	-	-	0.40	-	-	0.41	
MARBERT	-	-	0.29	-	-	-	-	-	-	-	-	-	-
RoBERTa-BNE	-	-	-	-	-	-	0.66	-	-	-	-	-	-
HauRoBERTa	-	-	-	-	-	-	-	0.12	-	-	-	-	-
	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	0.40	0.80	0.62	0.62	0.76	0.47	0.57	0.84	0.82

Table 5: Spearman correlation scores for different fine-tuned models in the three settings that we describe (supervised, unsupervised, and crosslingual) in addition to a simple lexical overlap baseline (Overlap).

- sentence embeddings (LaBSE (Feng et al., 2020), SentenceBERT (Reimers and Gurevych, 2019)), and
- standard encoder-based embeddings.

4.3 Lexical Overlap

As shown in Table 5, we report a simple lexical overlap baseline which consists of the Dice coefficient between two sentences A and B: the number of unique unigrams occurring in both sentences, adjusted by their lengths (Abdalla et al., 2023):

$$\frac{2 \times |unigram(A) \cap unigram(B)|}{|unigram(A) + unigram(B)|} \quad (1)$$

4.4 Supervised and Crosslingual settings

We use LaBSE (Label Agnostic BERT Sentence Embeddings) (Feng et al., 2020) which can map 109 languages into a shared vector space. With the embeddings covering all the SemRel languages, we report baseline results using the default hyperparameters set in the sentence-transformers repository[4]. Our experiments are conducted:

- using the predefined setup without further fine-tuning,
- by fine-tuning the LaBSE model on our training data using a cosine similarity loss.

We report the scores on the test sets in both setups in Appendix A, Table 12.

For the crosslingual baselines, we fine-tune LaBSE on the English training set and test on all the other datasets except English. On the other hand, when testing on the English dataset, we use the Spanish training set to fine-tune LaBSE.

4.5 Unsupervised settings

We used the standard encoder-based monolingual and multilingual language models on our datasets[5]. We experiment with:

- multilingual BERT (mBERT) (Devlin et al., 2019), XLMRoberta (XLMR) (Conneau et al., 2020) for all 13 languages,
- monolingual models:
  - AfroXLMR (Alabi et al., 2022) for Afrikaans, Amharic, Hausa, Kinyarwanda,
  - Indic-BERT (Kakwani et al., 2020) for Hindi, Marathi, and Telugu,
  - BERT (Devlin et al., 2019) for English,
  - MARBERT, ARBERT (Abdul-Mageed et al., 2021) and Arabic BERT (Safaya et al., 2020) for Arabic,
  - BETO (Cañete et al., 2020), ALBETO (Cañete et al., 2022), and RoBERTa-

[4] https://github.com/UKPLab/used-transformers

[5] We use the standard models from HuggingFace (Wolf et al., 2020).

BNE (Fandiño et al., 2022) for Spanish,
- Amharic RoBERTa (AmRoBERTa) (Yimam et al., 2021) for Amharic,
- DziriBERT (Abdaoui et al., 2021) for Algerian Arabic,
- RoBERTa based model (HauRoBERTa) for Hausa (Adelani et al., 2022).

We report the Spearman correlation scores with cosine similarity scores for all the BERT-based models in Table 5. Additional results using BERTScore (Zhang* et al., 2020) for mBERT and XLMR are shared in the Appendix (see Table 13).

4.6 Experimental results

Table 5 shows the Spearman correlation scores for the three setups: supervised, unsupervised, and crosslingual for all thirteen languages. For the unsupervised models, we report the results using all pretrained models including mBERT (Devlin et al., 2019) and XLMR (Conneau et al., 2020). Additionally, we report the Spearman correlation for experiments with monolingual language-specific models for each language – models that have been trained in these specific languages.

For the general setup, we note that, except for Amharic and Kinyarwanda, mBERT outperforms XLMR in all languages by a significant margin. For Amharic, mBERT's correlation score with the gold labels is 0.13, whereas XLMR's is three times higher with a 0.57 correlation score. Surprisingly, even though Arabic is a high-resource language, the Spearman correlation score is relatively low in comparison to all the high-resource languages, with Spanish achieving the best results. This could be due to the size of the Arabic data being smaller.

For the language-specific models, the results are highly tied to the language. In cases such as Amharic for example, AmRoBERTa significantly improves the score by 0.27 points, whereas AfroXLMR hurts the performance for all African languages.

Similarly to the unsupervised setup, high-resource languages have the highest scores in supervised and crosslingual settings. Overall, we report relatively higher correlation scores which vary between 0.40 and 0.88.

5 Conclusion

We presented SemRel, a new collection of semantic textual relatedness datasets in 13 languages with the majority predominantly spoken in Africa and Asia and considered low-resource. The sentence pairs contained in the datasets are annotated by native speakers and are associated with fine-grained relatedness scores. We reported the details related to the data curation and emphasised the challenges faced when dealing with low-resource languages.

We publicly release the datasets as well as other resources, such as the annotation guidelines and full labels for the research community interested in semantic relatedness, low-resource languages, and disagreements.

6 Limitations

We acknowledge that there is no formal definition of what constitutes semantic relatedness. Hence, the annotations may be subjective. To mitigate the issue we share our guidelines and annotated instances so researchers in the community can expand on our work, replicate, and study the disagreements in our data. We are also aware of the limited number of data sources and data variety in some low-resource languages involved. We do not claim that the datasets released represent all variations of these languages but they remain a good starting point as they were carefully picked, labelled, and processed by native speakers.

Although our collection is comprised of multiple datasets, the size of the data is limited, thus it cannot be the only source used for tasks that require a large amount of data such as language identification.

7 Ethical Considerations

All the annotators involved in this study were either volunteers or were paid more than the minimum wage per hour and any demographic information reported in the Appendix was shared with consent. The data that was further annotated was publicly available and is cited in our paper.

Similarly to Abdalla et al. (2023), we acknowledge all the possible socio-cultural biases that can come with our data, due to the data sources or the annotation process. When building our datasets, we did avoid instances with inappropriate or offensive utterances but we might have missed some. Our goal was to identify common perceptions of semantic relatedness by native speakers and our labels are not meant to be standardised for any given language. Note that we build datasets for low-resource languages but we do not claim in any way that these are fully representative of their usage.

Acknowledgements

We thank our annotators for labelling the data and for the insightful comments as well as Zara Siddique for providing additional insights.

Thanks to Dimosthenis Antypas, Joanne Boisson, Hsuvas Borkakoty for the helpful feedback.

References

Miguel Á. García Cumbreras, L. Alfonso Ureña López, and Fernando Martínez Santiago. 2006. BRUJA: Question classification for Spanish. using machine translationand an English classifier. In *Proceedings of the Workshop on Multilingual Question Answering - MLQA '06*.

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif Mohammad. 2023. What makes sentences semantically related? a textual relatedness dataset and empirical study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 782–796, Dubrovnik, Croatia. Association for Computational Linguistics.

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2021. Dziribert: A pre-trained language model for the Algerian dialect. *arXiv preprint arXiv:2109.12346*.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Idris Abdulmumin and Bashir Shehu Galadanci. 2019. hauwe: Hausa words embedding for natural language processing. In *2019 2nd International Conference of the IEEE Nigeria Computer Chapter (NigeriaComputConf)*. IEEE.

Karima Abidi, Mohamed Amine Menacer, and Kamel Smaili. 2017. CALYOU: A comparable spoken Algerian corpus harvested from YouTube. In *18th Annual Conference of the International Communication Association (Interspeech)*.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistic.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics).

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013a. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013b. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43.

Mahtab Ahmed, Chahna Dixit, Robert E Mercer, Atif Khan, Muhammad Rifayat Samee, and Felipe Urra. 2020. Multilingual corpus creation for multilingual semantic similarity task. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4190–4196.

Ahmed Hashim Al-Dulaimi. 2022. *Ultimate Arabic News Dataset*.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Vladimir Araujo, Andrés Carvallo, Souvik Kundu, José Cañete, Marcelo Mendoza, Robert E. Mercer, Felipe Bravo-Marquez, Marie-Francine Moens, and Alvaro Soto. 2022. Evaluation benchmarks for Spanish sentence representations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6024–6034, Marseille, France. European Language Resources Association.

Carlos Santos Armendariz, Matthew Purver, Senja Pollak, Nikola Ljubešić, Matej Ulčar, Ivan Vulić, and Mohammad Taher Pilehvar. 2020. SemEval-2020 task 3: Graded word similarity in context. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 36–49, Barcelona (online). International Committee for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

José Cañete, Sebastian Donoso, Felipe Bravo-Marquez, Andrés Carvallo, and Vladimir Araujo. 2022. ALBETO and DistilBETO: Lightweight Spanish language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4291–4298, Marseille, France. European Language Resources Association.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017a. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017b. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Asier Gutiérrez Fandiño, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852*.

Charles A Ferguson. 1959. Diglossia. *word*, 15(2):325–340.

Terry N Flynn and Anthony AJ Marley. 2014. *Best-worst scaling: Theory and methods*. Ph.D. thesis, Edward Elgar Worcester, UK.

Goran Glavaš, Marc Franco-Salvador, Simone P Ponzetto, and Paolo Rosso. 2018. A resource-light method for cross-lingual semantic textual similarity. *Knowledge-based systems*, 143:1–9.

Ruqaiya Hasan and Michael AK Halliday. 1976. Cohesion in English. *London, 1976; Martin JR*.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*.

Javier Huertas-Tato, Alejandro Martín, and David Camacho. 2021. Silt: Efficient transformer training for inter-lingual inference.

Abderrahmane Issam and Khalil Mrini. 2022. Goud.ma: a news article dataset for summarization in Moroccan Darija. In *3rd Workshop on African Natural Language Processing*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817, San Diego, California. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. *arXiv preprint arXiv:1712.01765*.

G Frederic Kuder and Marion W Richardson. 1937. The theory of the estimation of test reliability. *Psychometrika*, 2(3):151–160.

Kemal Kurniawan and Samuel Louvan. 2018. Indosum: A new benchmark dataset for Indonesian text summarization. In *2018 International Conference on Asian Language Processing (IALP)*, pages 215–220. IEEE.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Jordan J Louviere and George G Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. Technical report, Working paper.

Lokesh Madasu, Gopichand Kanumolu, Nirmal Surange, and Manish Shrivastava. 2023. Mukhyansh: A headline generation dataset for Indic languages. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 620–634, Hong Kong, China. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2014)*, pages 216–223. Reykjavik.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on PADIC: A parallel Arabic dialect corpus. In *Proceedings of the 29th Pacific Asia conference on language, information and computation*, pages 26–34.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Saif Mohammad. 2008. *Measuring Semantic Distance Using Distributional Profiles of Concepts*. Ph.D. thesis, University of Toronto.

Saif M Mohammad and Graeme Hirst. 2012. Distributional Measures of Semantic Distance: A Survey. *arXiv preprint arXiv:1203.1858*.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1):21–48.

Rubungo Andre Niyongabo, Qu Hong, Julia Kreutzer, and Li Huang. 2020. KINNEWS and KIRNEWS: Benchmarking cross-lingual text classification for Kinyarwanda and Kirundi. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5507–5521, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bryan K. Orme. 2009. Maxdiff analysis : Simple counting , individual-level logit , and hb.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Rajat Pandit, Saptarshi Sengupta, Sudip Kumar Naskar, Niladri Sekhar Dash, and Mohini Mohan Sardar.

2019. Improving semantic similarity with cross-lingual resources: A study in Bangla—a low resourced language. In *Informatics*, volume 6, page 19. MDPI.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Ken Nabila Setya and Rahmad Mahendra. 2018. Semi-supervised textual entailment on Indonesian wikipedia data. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 416–427. Springer.

Xin Tang, Shanbo Cheng, Loc Do, Zhiyu Min, Feng Ji, Heng Yu, Ji Zhang, and Haiqin Chen. 2018. Improving multilingual semantic textual similarity with shared sentence encoder for low-resource languages. *arXiv preprint arXiv:1810.08740*.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, et al. 2020. Multi-simlex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.

Mengting Wan and Julian McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM conference on recommender systems*, pages 86–94.

Bin Wang, C.-C. Jay Kuo, and Haizhou Li. 2022. Just rank: Rethinking evaluation with word and sentence similarities. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6060–6077, Dublin, Ireland. Association for Computational Linguistics.

Bin Wang and Haizhou Li. 2022. Relational sentence embedding for flexible semantic matching. *arXiv preprint arXiv:2212.08802*.

John Wieting and Kevin Gimpel. 2017. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *arXiv preprint arXiv:1711.05732*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Seid Muhie Yimam, Abinew Ali Ayele, Gopalakrishnan Venkatesh, Ibrahim Gashaw, and Chris Biemann. 2021. Introducing various semantic models for Amharic: Experimentation and evaluation with multiple tasks and datasets. *Future Internet*, 13(11).

Yunjin Yum, Jeong Moon Lee, Moon Joung Jang, Yoojoong Kim, Jong-Ho Kim, Seongtae Kim, Unsub Shin, Sanghoun Song, and Hyung Joon Joo. 2021. A word pair dataset for semantic similarity and relatedness in Korean medical vocabulary: Reference development and validation. *JMIR Medical Informatics*, 9(6):e29667.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Chengqing Zong. 2015. Improving SMT by model filtering and phrase embedding. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Keynotes*, Da Nang, Vietnam.

Appendix

A Annotation

A.1 Pilot data annotation

To assess the different pairing techniques and the potential annotation challenges, we run a pilot annotation task on 20 to 100 pairs of sentences for each language before proceeding with larger annotation batches. This helped us assess the difficulties related to the annotation task and the choices to be

Lang.	Datasets (%)
afr	Oscar (100%).
amh	News data (100%).
arb	Ted Talk subtitles on science, art, and society (96%), Economy news data (4%).
arq	Conversational data (89%), Youtube comments (11%).
ary	News data (100%).
eng	Wikipedia (29%), ParaNMT (17%), Formality (17%), SNLI (8%), Goodreads (22%), STS (7%).
esp	MuchoChine (MC) (7%), Spanish QC (4%), PAWS-X (19.5%), NLI-es (12%), SICK-es (17%), STS (5%), BSO (17.5%), XL-Sum (18%).
hau	News data (100%)
hin	News data (100%).
ind	News data (82%), Wikipedia/ROOTS (18%).
kin	News data (100%).
mar	News data (100%).
tel	News data (100%).

Table 6: The percentage of instances collected from different sources (datasets).

made for the final data processing step. For instance, if highly related and unrelated pairs were occurring too often in the tuples, we reduced the percentages of both highly related and unrelated pairs by changing or calibrating the data sources if possible, prioritising other pairing techniques, or including an extra preprocessing step (e.g., paraphrase detection).

A.2 Data Pre-processing Tools

We used NLTK tools for parsing Afrikaans and Indonesian in addition to manual verification. For instance, as for Indonesian, the NLTK sentence parses generated many errors due to common Indonesian abbreviations that involve *'.'*, which the sentence parser mistakenly detects as the end of a sentence, we added new abbreviations for parsing ('ir.', 'kh.', 'h.', 'drs.', 'drg.', 'rm.', 'bp.', 'bpk.', 'tgl.', 'no.', 'jl.', and 'jln.')

A.3 Information about the Annotators

We report on the demographic information of the volunteers who agreed to share them.

Afrikaans Paid native speakers.

Amharic Paid Amharic native speakers, 3 women and 5 men from different social, cultural, and ethnic backgrounds (Amhara, Guragie, Wolyta, Sidama, and Oromo).

Modern Standard Arabic and Algerian Arabic Native speakers, 2 men, 2 women, university degree holders, ages vary between 23 to 56, paid above the minimum wage.

Moroccan Arabic Volunteer native speakers, 3 women, 1 man, university degree holders, volunteers.

English and Spanish Amazon Mechanical Turkers with high approval rates (98% for English) paid above the US minimum wage.

Hausa Paid native speakers, 3 women, 1 man, age: 28 to 30, bachelor's degree holders.

Hindi and Marathi Paid native speakers.

Telugu Volunteer native speakers.

A.4 Annotation Guidelines

You will be given four sentence pairs (i.e., 4 pairs of the form [sentence A, sentence B]). Your task is to judge the relatedness of each pair (sentence A and sentence B) and **tell us**:

- the sentence pair that is the **MOST related** (i.e., sentence A is closest in meaning to sentence B).
- the sentence pair that is the **LEAST related** (i.e., sentence A is farthest in meaning to sentence B).

Sentence pairs can be related in many ways. I.e., sentence A and sentence B can be related in different ways. The first pair of sentences in Table 7 are more related than the second one. Often, sentence pairs that are more specific in what they share tend to be more related than sentence pairs that are only loosely about the same topic.

If a sentence has more than one interpretation, consider that meaning which is closest to the meaning of the other sentence in the pair. If both sentences have multiple meanings, then consider those meanings that are closest to each other.

If in the given set of four pairs, two (or more) sentence pairs are **equally related** to each other

MOST Related Pair	S1: The boy enjoyed reading under the lemon tree S2: There is a lemon tree next to the house
LEAST related Pair	S1: The boy enjoyed reading under the lemon tree S2: The boy was an excellent football player

Table 7: Example in the Guidelines. Examples of two pairs of sentences with different degrees of relatedness from Abdalla et al. (2023).

Pair 1	S1: The boy enjoyed reading under the lemon tree S2: I have a green hat
Pair 2	S1: The boy enjoyed reading under the lemon tree S2: She was an excellent football player

Table 8: Example in the Guidelines. Examples of two pairs of sentences that have similar degrees of relatedness where one can choose randomly the most vs. least related pairs (i.e., either Pair 1 or Pair 2).

and they are also the most related pairs, then select **either** one of them as the most related (i.e., **randomly**). Similarly, if two (or more) equally related pairs are also the least related pairs, then select either one of them as the least related. (See Table 2.)

You cannot select the same sentence pair for both categories.

Try not to overthink the answer. Let your instinct guide you.

A.5 Notes

Sentence pairs can be related in many ways. Consider the entire meaning of the sentences before selecting the most related. The sentences included in this task may contain foul language, though we have attempted to limit this.

A.6 Examples (Q1)

Which of the four sentence pairs in Table 9 is MOST RELATED? Which pair is LEAST RELATED?

A.6.1 A1

The most related pair is Pair 3 because both sentences are talking about a group sitting/resting in grass.

The least related is Pair 4 because Pair 4 sentences are completely unrelated, whereas the other pairs have some relatedness.

A.6.2 Note (A1)

Pair 1 sentences are somewhat related, as they talk about Narnia/characters in that world (Aslan and Bree are characters in Narnia). However, the content of this sentence pair is not as related as Pair 3.

Pair 2 sentences are both talking about romantic relationships.

A.7 Examples (Q2)

Which of the four sentence pairs in Table 10 is MOST RELATED? Which pair is LEAST RELATED?

A.7.1 A2

The most related pair is Pair 4. Both sentences are talking about the same city and mention that it is on the bank of river Sarayu. **The least related pair** is Pair 2 because the sentences are completely unrelated.

A.7.2 Note (A2)

Pair 3 sentences both refer to at least one woman outside.

Pair 1 sentences refer to kids or kid-related things (making them slightly close in meaning).

A.8 Examples (Q3)

Which of the four sentence pairs in Table 11 is MOST RELATED? Which pair is LEAST RELATED?

A.8.1 A3

The most related pair is Pair 4. Both sentences are paraphrases of each other. (Pair 1 and Pair 2 are quite related but not as exact paraphrases as Pair 4.)

The least related pair is Pair 3. Pair 3 sentences are somewhat related as they talk about house furnishings. However, they are still less related than all the other pairs.

Pair 1	S1: My personal favorites from Narnia were the conversations between Aslan and Bree. S2: This marks my progress through the Chronicles, picked up after reading The Narnia Code and Planet Narnia.
Pair 2	S1: why won't she ask me out? S2: and after all that you wont have to worry about getting a girl to like you.
Pair 3	S1: A group of people are sitting on the grass outside of a rustic building. S2: Group sitting on a grassy hill resting.
Pair 4	S1: If you change me back, I will feed each one of your snakes a large mouse! S2: Offer people who join cash and coupons.

Table 9: Q1 Example in the Guidelines.

Pair 1	S1: That and a kids meal. S2: My two kids, ages 5 and 3!
Pair 2	S1: The spines , which may be up to 50 mm long , are modified hairs , mostly made of keratin . S2: The simplest shape is the long opening with a pointed arch known in England as the lancet .
Pair 3	S1: A woman wearing a white shirt and a red headband is sitting outside. S2: Two women stand outside a library.
Pair 4	S1: Ayodhya ,capital of King Rama is mentioned on the banks of Sarayu river . S2: Ramayana mentions that city of Ayodhya was situated on the bank of Sarayu river .

Table 10: Q2 Example in the Guidelines.

Pair 1	S1: IBM has not shifted its focus from mainframes to compete with Windows S2: In 3 years, IBM has not been interested in the PC.
Pair 2	S1: I wanted to see the scene where Quinn told the brotherhood he was in love with Blay. S2: I also would have liked to see the scene where Qhuinn asks Blay's dad for permission to propose to Blay.
Pair 3	S1: Jeremy desperately needs a stable home. S2: Furnishings were an angle bed, a stool, and a chamber pot on the dirt floor.
Pair 4	S1: That's difficult. They're both great S2: that's really hard they are both great!

Table 11: Q3 Example in the Guidelines.

A.8.2 Note (A3)

Pair 1 sentences both refer to IBM and their business strategy. We consider this to be more related than Pair 3 because it's more specific in the details they share.

Pair 2 sentences talk about the same characters and their romantic situation.

B Pre-trained models used

We list down the various pre-trained HuggingFace models used in our experiments:

1. mBERT
2. XLMR
3. AfroXLMR
4. ALBETO
5. AmRoBERTa
6. ARBERT
7. `arb` BERT
8. BETO
9. DziriBERT
10. Indic-BERT
11. MARBERT
12. RoBERTa-BNE
13. HauRoBERTa
14. LaBSE

Language	Base	Finetuned
afr	0.76	0.79
amh	0.79	0.85
arb	0.55	0.62
arq	0.40	0.60
ary	0.38	0.77
eng	0.82	0.83
esp	0.65	0.70
hau	0.48	0.69
hin	0.71	0.77
ind	0.53	0.50
kin	0.45	0.72
mar	0.82	0.88
tel	0.80	0.82

Table 12: Spearman correlation scores on LaBSE models with and without further fine-tuning on our training data (Base and fine-tuned, respectively).

Language	mBERT	XLMR
afr	0.77	0.76
amh	0.12	0.69
arb	0.40	0.42
arq	0.28	0.32
ary	0.53	0.50
eng	0.71	0.74
esp	0.67	0.68
hau	0.32	0.31
hin	0.64	0.63
ind	0.54	0.54
kin	0.25	0.30
mar	0.78	0.75
tel	0.77	0.78

Table 13: Spearman correlation of the BERTScore (Zhang* et al., 2020) with mBERT and XLMR on the different languages.

Lang.	Sentence 1	Sentence 2	Score
tel	ఇప్పటికే ధోనీ టెస్టు క్రికెట్ కు గుడ్ బై చెప్పిన విషయం తెలిసిందే. Gloss: It is already known that Dhoni has said goodbye to Test cricket.	అంతకు ముందు జరిగిన మరో రోడ్డు ప్రమాదంలో ఏడుగురు మరణించగా, 12 మంది గాయపడ్డారు. Gloss: Seven people were killed and 12 injured in another road accident earlier.	0.02
afr	My eerste stukkie advies is dat jy realisties moet wees oor die afstand wat jy wil hengel. **Gloss:** My first piece of advice is to be realistic about the distance you want to fish.	Dit bring tot n einde die maanverkenningsprogram van die Verenigde State.. **Gloss:** This brings to an end the lunar exploration program of the United States.	0.19
esp	Costo monetario para mantener el microondas por $6. **Gloss:** Monetary cost to maintain the microwave is $6.	Todavía nos quedan más de 200.000$ por recaudar de suscriptores y donantes como usted. **Gloss:** We still have over $200,000 left to raise from subscribers and donors like you.	0.27
mar	ठाकरे सरकारच्या मंत्रिमंडळात 25 कॅबिनेट मंत्री असणार आहेत. **Gloss:** Thackeray government will have 25 cabinet ministers.	त्यामुळे गुढी पाडवा मेळाव्यामध्ये राज ठाकरे काय बोलणार याकडे सर्वांचे लक्ष लागून राहिले आहे. **Gloss:** Therefore, everyone's attention is on what Raj Thackeray will say in the Gudi Padwa gathering..	0.42
arq	كاين واحد الأبيات يقولهم في الغنى تاعو تكوني تعرفيهم **Gloss:** There's a couplet in one of his songs that you may know.	اللي ما زهاش في الدنيا من الروح خالي **Gloss:** "He who did not feel joy in this world is empty-spirited/has no soul" [a couplet]	0.50
arb	الآن انتقل بكم إلى لمحة عن تاريخ الاقتصاد و في نظري قد تكون مفيدة. **Gloss:** Now, I will take you to a glimpse of the history of economics, which in my opinion may be useful.	حوالي عام ١٨٥٠ كان صيد الحيتان من أكبر الصناعات في الولايات المتحدة **Gloss:** Around 1850, whaling was one of the largest industries in the United States.	0.62
hin	देश में कोरोना वायरस से मौत का आंकड़ा 100 के पार पहुंचा, पिछले 12 घंटे में 26 की गई जान. **Gloss:** Death toll due to Corona virus in the country crossed 100, 26 people lost their lives in the last 12 hours.	देश में कोरोना वायरस का कहर तेजी से बढ़ता जा रहा है। **Gloss:** The havoc of Corona virus is increasing rapidly in the country.	0.72

Continued on next page

Table 13 – continued from previous page

Lang.	Sentence 1	Sentence 2	Score
kin	Duhugukire kwandika neza Ikinyarwanda Mu myandikire y'Ikinyarwanda hari amakosa akunda gukorwa ashingiye ku ifatana n'itandukana ry'amagambo. **Gloss:** Let's learn to write well in Ikinyarwanda In writing Ikinyarwanda there are mistakes that are often made based on the connection and separation of words.	Duhugukire kwandika neza Ikinyarwanda (igice cya gatatu) Mu myandikire y'Ikinyarwanda, hari amagambo afatana n'andi atandukana. **Gloss:** Let's practice writing well in Ikinyarwanda (part three) In Ikinyarwanda writing, there are words that go together and others that are different.	0.75
ary	وجدو راسكوم لرمضان.. الحرارة غادي تبدا بـ٣٧ درجة فهاد المناطق **Gloss:** Prepare yourselves for Ramadan, the temperature will start at 37 degrees in these regions.	غير خرج رمضان وهي تشعل.. الحرارة غادي تبدا وغادي توصل لـ٤٠ درجة فهاد المناطق **Gloss:** Since Ramadan, the weather has been very hot. Temperatures are rising and they will reach 40 degrees in these regions.	0.75
ind	Pendidikan Desa Pusaka memiliki 4 sekolah. **Gloss:** Pusaka Village Education has 4 schools.	Pendidikan Desa Serumpun Buluh memiliki 4 sekolah. **Gloss:** Serumpun Buluh Village Education has 4 schools.	0.83
amh	እኛን ከዚህ ጉዳይ ጋር የምንገናኝበት ቅንጣት ታክል ግንኙነት የለም **Gloss:** There is nothing of a connection that concern us with this issue.	እኛን ቅንጣት ታህል ከዚህ ጉዳይ ጋ የሚያገናኘን ነገር የለም **Gloss:** There is nothing that concern us with this issue.	0.89
hau	Haka ya furta a cikin jawabin sa na murnar cikar Najeriya shekaru 61 da samun 'yanci. **Gloss:** That is what he said in the speech for celebrating Nigeria's 61 independence day celebration.	Ya yi wannan ikirarin ne a cikin jawabin sa na murnar cikar Najeriya 61 da samun 'yanci a ranar Juma'a. **Gloss:** He made this assertion in his speech celebrating Nigeria's 61 independence day celebration on Friday.	0.94
eng	I've been searching the entire abbey for you.	I'm looking for you all over the abbey.	1.00

Table 13: Examples of SemRel data instances and their translations.