SemEval-2024 Task 1: Semantic Textual Relatedness for African and Asian 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], Meriem Beloucif[8], Christine De Kock[9], Oumaima Hourrane, Manish Shrivastava[10], Thamar Solorio[6], Nirmal Surange[10], Krishnapriya Vishnubhotla[11], Seid Muhie Yimam[12], Saif M. Mohammad[13]

[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]Uppsala University, [9]The University of Merlbourne, [10]IIIT Hyderabad, [11]University of Toronto, [12]Universität Hamburg, [13]National Research Council Canada.

Contact: OusidhoumN@cardiff.ac.uk, s.muhammad@imperial.ac.uk

Abstract

We present the first shared task on Semantic Textual Relatedness (STR). While earlier shared tasks primarily focused on semantic similarity, we instead investigate the broader phenomenon of semantic relatedness across 14 languages: *Afrikaans, Algerian Arabic, Amharic, English, Hausa, Hindi, Indonesian, Kinyarwanda, Marathi, Moroccan Arabic, Modern Standard Arabic, Punjabi, Spanish,* and *Telugu*. These languages originate from five distinct language families and are predominantly spoken in Africa and Asia – regions characterised by the relatively limited availability of NLP resources. Each instance in the datasets is a sentence pair associated with a score that represents the degree of semantic textual relatedness between the two sentences. Participating systems were asked to rank sentence pairs by their closeness in meaning (i.e., their degree of semantic relatedness) in the 14 languages in three main tracks: (a) supervised, (b) unsupervised, and (c) crosslingual. The task attracted 163 participants. We received 70 submissions in total (across all tasks) from 51 different teams, and 38 system description papers. We report on the best-performing systems as well as the most common and the most effective approaches for the three different tracks.

1 Introduction

Defining 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 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). For example, *'I am feeling sick.'* and *'Get well soon!'* would receive a low similarity score, despite the two being very related. In this shared task, we investigate the broader concept of semantic textual relatedness. 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 shared tasks (Agirre et al., 2012, 2013, 2014, 2015, 2016; Cer et al., 2017) have mainly focused on textual similarity. In this work, we provide participants with SemRel (Ousidhoum et al., 2024), a collection of 14 newly curated monolingual STR datasets 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`), Punjabi (`pun`) and Telugu (`tel`). The datasets are composed of sentence pairs, each assigned a relatedness score between 0 (completely

*Equal contribution from first and second authors, authors 3 to 16 are alphabetically ordered.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1963–1978
June 20-21, 2024 ©2024 Association for Computational Linguistics

Lang.	Family	Train	Dev	Test
afr	Indo-European	-	375	375
amh	Afro-Asiatic	992	95	171
arb	Afro-Asiatic	-	32	595
arq	Afro-Asiatic	1,261	97	583
ary	Afro-Asiatic	925	70	427
eng	Indo-European	5,500	250	2,600
esp	Indo-European	1,562	140	600
hau	Afro-Asiatic	1,763	212	603
hin	Indo-European	-	288	968
ind	Austronesian	-	144	360
kin	Niger-Congo	778	102	222
mar	Indo-European	1,200	293	298
pan	Indo-European	-	638	242
tel	Dravidian	1,170	130	297

Table 1: The language families and data split sizes of the different datasets. Datasets with no traning sets were only used in tracks B and C.

unrelated) and 1 (maximally related) with a large range of expected relatedness values. 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. 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; Kiritchenko and Mohammad, 2017a). The shared task included three main tracks: (1) supervised, (2) unsupervised, and (3) cross-lingual.

Each team could provide submissions for one, two, or all of the tracks in one or more languages. Our official evaluation metric was the Spearman rank correlation coefficient, which captures how well the system-predicted rankings of test instances aligned with human judgments. Our task attracted 163 participants, received 70 final submissions from 51 different teams, and 38 teams submitted system description papers. Track A (supervised) received the largest number of submissions: 40, followed by 18 submissions for track B (unsupervised) and 12 for track C (crosslingual). Most teams participated in multiple languages (more than eight on average). All of the task details and resources are available on the task website.[1]

2 Related Work

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

Most prior shared tasks focus on semantic textual similarity, a narrower subset of relatedness, and often only cover high-resource languages such as English (Agirre et al., 2012, 2013, 2014, 2015, 2016), Arabic, German, Spanish, and Turkish (Cer et al., 2017) with few exceptions such as Armendariz et al. (2020) who also included Slovene, Finnish, and Croatian.

By comparison, this shared task focuses on sentence-level STR in various low-resource languages. 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. The core of Abdalla et al. (2023) approach served as the model for data annotations added to new ways of data collection–curation for several less-resourced languages.

3 Data

3.1 Data Collection

A key step in the data creation process was identifying text sources for each language and selecting sentence pairs. This was particularly challenging for low-resource languages such as Hausa, Telugu, or Algerian Arabic. Since most 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. Thus, data instances were extracted for each language using a tailored combination of heuristics such as lexical overlap and paraphrases. We used further pre-processing, post-processing, and data analysis methods to avoid incoherence and unnaturalness.

Since arbitrarily selecting sentences and pairing them would lead to many unrelated instances, we relied on the following heuristics to pair sentences and ensure that the pairs would exhibit relatedness scores varying from completely unrelated to very related:

1. **Lexical Overlap** Select sentences with various proportions of lexical overlap, i.e., one or more words/tokens in common, with or without using TF/IDF normalisation.
2. **Contiguity/Entailment** Select adjacent pairs of sentences in a paragraph or a social media

[1] https://semantic-textual-relatedness.github.io

Language	afr	amh	arb	arq	ary	eng	esp	hau	hin	ind	kin	mar	pun	tel
#Annotators	2	4	2-3	2	2	2-4	2-4	2-4	4	2	2	2-3	2	4
SHR train/dev	0.85	0.89	0.86	0.64	0.77	0.80	0.70	0.74	0.93	0.68	0.74	0.92	0.65	0.79
SHR test	0.85	0.89	0.86	0.64	0.77	0.80	0.70	0.74	0.93	0.68	0.74	0.92	0.65	0.79

Table 2: SHR (split-half reliability) scores for each of the created dataset splits and numbers of annotators per tuple (#Annotators).

thread, i.e., sentences that appear one after the other.

3. **Paraphrases or Machine Translation (MT) Paraphrases** Select 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.
4. **Random selection** Random pairs of sentences are selected.

We elaborate on the detailed data collection and processing steps in Ousidhoum et al. (2024).

3.2 Data Annotation

As the notions of *related* and *unrelated* do not have clear boundaries with no unanimous definition in the literature, we use comparative annotations and rely on the intuitions of fluent speakers for each language to choose between sentence pairs. Therefore, instead of relying on vague class definitions, we capture common perceptions of semantic relatedness (i.e., what is believed by the vast majority) rather than "correct" or "right" rankings.

We used Best–Worst Scaling (BWS) (Louviere and Woodworth, 1991; Kiritchenko and Mohammad, 2017a), a form of comparative annotation that avoids various biases of traditional rating scales, to annotate our data instances and generate an ordinal ranking of instances. In BWS, annotators are given n items (an n-tuple, where $n > 1$ and commonly $n = 4$). They are asked which item is the *best* (highest in terms of the property of interest) and which is the *worst* (lowest in terms of the property of interest). When working on 4-tuples, best–worst annotations are particularly efficient because each best and worst annotation will reveal the order of five of the six-item pairs. Real-valued scores of association between the items and the property of interest can be determined using simple arithmetic on the number of times an item was chosen best and the number of times it was chosen worst (Orme, 2009; Flynn and Marley, 2014). It has been empirically shown that annotations for $2N$ 4-tuples are sufficient for obtaining reliable scores (where N is the number of items) (Louviere and Woodworth, 1991; Kiritchenko and Mohammad, 2016). Kiritchenko and Mohammad (2017b) showed through empirical experiments that BWS produces more reliable and discriminating scores than those obtained using rating scales. (See (Kiritchenko and Mohammad, 2016, 2017b) for further details on BWS.) We generated tuples using the BWS scripts provided by Kiritchenko and Mohammad (2017a)[2].

We report the number of annotators and the split-half reliability (SHR) scores (Cronbach, 1951; Kuder and Richardson, 1937) for each of the datasets in Table 2. SHR measures the degree to which repeating the annotations results in similar relative rankings of the instancesOverall the scores in Table 2 vary between 0.64 and 0.96, which indicates a high annotation reliability.

4 Task Description

In this task, we aim to predict the semantic textual relatedness (STR) of sentence pairs. Participants had to rank sentence pairs by their degree of semantic relatedness which varies between 0 (unrelated) and 1 (closely related). Each team could provide submissions for one, two, or all of the tracks presented below.

4.1 Track A: Supervised

Participants were to submit systems trained on the labeled training datasets provided. Participating teams were allowed to use any publicly available datasets (e.g., other relatedness and similarity datasets or datasets in any other languages). However, they had to report on additional data they used, and ideally report how each resource impacted the final results.

4.2 Track B: Unsupervised

Participants were to submit systems that were developed without the use of any labeled datasets

[2] https://saifmohammad.com/WebPages/BestWorst.html

#	Track A (Supervised)		Track B (Unsupervised)			Track C (Crosslingual)		
	Team	**Score**		**Team**	**Score**		**Team**	**Score**
			*	Lexical Overlap	0.456			
*	baseline (LaBSE)	0.762	*	baseline (XLMR)	0.353	*	baseline (LaBSE)	0.579
1	AAdam	0.800		SATLab	0.543		AAdaM	0.650
2	NRK	0.781		MasonTigers	0.514		UAlberta	0.589
3	PEAR	0.758		HW–TSC	0.482		silp_nlp	0.566
4	silp_nlp	0.740		UAlberta	0.481		MaiNLP	0.499
5	NLP_1@SSN	0.740		silp_nlp	0.400		ustcctsu	0.445

Table 3: Top 5 submissions per track. See Appendix for paper information about the different teams. * shows baseline results using lexical overlap, XLMR and LaBSE reported in the SemRel dataset paper (Ousidhoum et al., 2024).

pertaining to semantic relatedness or semantic similarity between units of text more than two words long in any language. The use of unigram or bigram relatedness datasets (from any language) was permitted.

4.3 Track C: Cross-lingual

Participants were to submit systems that were developed without the use of any labeled semantic similarity or semantic relatedness datasets in the target language and with the use of labeled dataset(s) from at least one other language. Using labeled data from another track was mandatory for a submission to this track.

4.4 Official Evaluation Metric

The official evaluation metric for this task is the Spearman rank correlation coefficient, which captures how well the system-predicted rankings of test instances align with human judgments. We provided the participants with an evaluation script on GitHub page[3].

4.5 Task Organisation

We released some pilot datasets before the start of the shared task for participants to have a better understanding of the task (i.e., the datasets, the languages involved, and the labels) and provided the participants with a starter kit on GitHub.

5 Evaluation

5.1 Our baselines

In Table 3, 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):

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

In addition, we used 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]. We used:

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

For the crosslingual baselines, we fine-tuned LaBSE on the English training set and tested on all the other datasets except English while using the Spanish training set to fine-tune LaBSE when testing on English. We elaborate on the detailed baseline experiment in (Ousidhoum et al., 2024)

5.2 Participating Systems and Results

5.3 Participant Overview

During the evaluation phase, 163 people registered for the competition. Of these, 51 teams made 70 final submissions across tracks [5]. Track A received 40 final submissions, track B received 12 submissions, and track C received 18. For track A, most participants submitted systems for at least eight languages. We report the top–5 performing systems in all tracks in Table 3.

[3]https://github.com/semantic-textual-relatedness/Semantic_Relatedness_SemEval2024

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

[5]The details can be found in the Appendix.

Rank	Team	amh	arq	ary	eng	esp	hau	kin	mar	tel	Average
1	AAdaM (Zhang et al., 2024)	0.867	0.662	0.835	0.848	**0.740**	0.724	0.779	0.894	0.848	0.800
2	NRK (Nguyen and Thin, 2024)	0.864	0.674	0.827	0.833	0.690	0.672	0.757	0.879	0.834	0.781
*	SemRel baseline (LaBSE)	0.789	0.847	0.761	0.830	0.702	0.693	0.725	0.881	0.817	0.762
3	PEAR (Jørgensen, 2024)	0.834	0.463	0.815	0.848	0.710	0.694	0.772	0.856	0.827	0.758
4	silp_nlp (Singh et al., 2024)	0.837	0.594	0.808	0.845	0.658	0.724	0.485	0.863	0.843	0.740
5	NLP_1@SSN (B et al., 2024)	-	0.623	0.745	0.835	0.705	0.628	0.723	0.871	0.789	0.740
6	UAlberta (Shi et al., 2024)	0.854	0.464	0.497	0.853	0.705	0.735	0.641	0.890	0.857	0.722
7	MBZUAI-UNAM (Ortiz-Barajas et al., 2024)	0.840	0.541	0.786	0.832	0.697	0.670	0.458	0.867	0.785	0.720
8	INGEOTEC (Moctezuma et al., 2024)	0.702	0.566	0.811	0.809	0.678	0.576	0.630	0.784	0.801	0.706
9	HausaNLP (Salahudeen et al., 2024)	0.353	0.587	0.834	0.794	0.723	0.594	0.633	0.837	0.800	0.684
10	KINLP	-	0.471	0.779	0.740	0.581	0.616	0.763	0.749	0.754	0.682
11	BITS Pilani (Venkatesh and Raman, 2024)	0.800	0.510	0.444	0.832	0.656	0.508	0.518	0.842	0.814	0.658
12	OZemi (Takahashi et al., 2024)	0.781	0.371	0.445	0.805	0.620	0.620	0.567	0.862	0.782	0.650
13	Text Mining (Keinan, 2024)	0.713	0.443	0.701	0.720	0.661	0.543	0.413	0.778	0.706	0.631
14	MasonTigers (Goswami et al., 2024)	0.785	0.400	0.376	0.836	0.651	0.477	0.367	0.818	0.802	0.612
15	YSP (Aali et al., 2024)	0.643	0.402	-	0.819	0.635	0.387	0.315	0.689	0.643	0.567
16	IITK (Basak et al., 2024)	0.550	0.339	0.358	0.808	0.591	0.219	0.138	0.666	0.282	0.439
17	YNUNLP2023 (Li et al., 2024b)	0.789	0.235	0.092	0.557	0.404	0.269	0.186	0.544	0.617	0.410
NR	PALI	**0.889**	0.679	**0.863**	**0.860**	0.724	**0.764**	0.813	**0.911**	0.864	**0.819**
NR	king001	0.888	**0.682**	0.860	0.843	0.721	0.747	**0.817**	0.897	0.853	0.812
NR	saturn	0.845	0.578	0.798	-	-	0.699	0.755	0.873	**0.873**	0.774
NR	UMBCLU (Roy Dipta and Vallurupalli, 2024)	-	-	0.745	0.838	0.721	0.640	0.681	0.841	0.682	0.733
NR	SemanticCUETSync (Hossain et al., 2024)	-	-	-	0.822	0.677	-	-	0.870	0.820	0.796
NR	NLP-LISAC (Benlahbib et al., 2024)	-	0.604	0.789	0.835	0.717	-	-	-	-	0.736
NR	Unknown	-	-	-	0.831	-	-	-	0.882	0.841	0.852
NR	BpHigh	-	-	-	0.809	-	-	-	0.875	0.769	0.819
NR	Sharif_STR (Ebrahimi et al., 2024)	-	0.380	-	0.827	0.673	-	-	-	-	0.441
NR	CAILMD-23 (Sonavane et al., 2024)	-	-	-	0.823	-	-	-	0.871	-	0.847
NR	WarwickNLP (Ebrahim and Joy, 2024)	-	-	0.816	0.842	-	-	-	-	-	0.829
NR	GIL-IIMAS UNAM	-	-	-	0.830	0.731	-	-	-	-	0.780
NR	msiino	-	-	-	0.809	0.611	-	-	-	-	0.710
NR	NLU-STR (Malaysha et al., 2024)	-	0.525	0.832	-	-	-	-	-	-	0.678
NR	Tübingen-CL (Zhang and Çöltekin, 2024)	-	-	-	0.850	-	-	-	-	-	0.850
NR	Pinealai (Eponon and Ramos Perez, 2024)	-	-	-	0.837	-	-	-	-	-	0.837
NR	gds142	-	-	-	-	-	-	-	-	0.826	0.826
NR	LuisRamos07	-	-	-	0.822	-	-	-	-	-	0.822
NR	VerbaNexAI Lab (Morillo et al., 2024)	-	-	-	0.819	-	-	-	-	-	0.819
NR	Fired_from_NLP (Shanto et al., 2024)	-	-	-	0.810	-	-	-	-	-	0.810
NR	Roronoa_Zoro	-	-	-	0.810	-	-	-	-	-	0.810
NR	NLP_STR_teamS (Su and Zhou, 2024)	-	-	-	0.809	-	-	-	-	-	0.809
NR	DataJo	-	0.356	-	-	-	-	-	-	-	0.356

Table 4: Track A results. The best results are in bold, and NR stands for *not ranked*. As the methods are highly language-dependent, we only rank teams that participated in at least 8 sub-tracks, but we highlight in blue the best results achieved by non-ranked teams. (Non-ranked teams are sorted based on the number of languages they participated in.)

5.4 Task A: Supervised

5.4.1 Best Performing Systems

AAdaM They opted for data augmentation by translating the English SemRel dataset and STSB (semantic similarity) to create and augment data in other languages. The team explored both fine-tuning and adapter-based tuning. Given a target language, they first fine-tuned the cross-encoder-based AfroXLMR model (Alabi et al., 2022) on the augmented data as a warm-up or TAPT (Task-Adaptive-Pre-Training) and then continued the fine-tuning on the provided SemRel data.

NRK They ensembled various BERT-like models and used a weighted voting technique to improve the performance of their model.

PEAR They examined the effect of combining or using per-language data through 5-fold validation. They did not conduct any text preprocessing to maintain fairness across languages. They defined three model configurations: "base" with no training, "all" trained on all languages, and "lan" trained on one language. They experimented with multilingual embeddings, cross-encoders, and augmented data from bi-encoders.

5.4.2 Popular Methods

The general trend for the methods submitted to track A was (1) embedding sentence pairs into text and (2) training a regression model. Some teams used traditional embeddings and regression approaches (e.g., word2vec with support vector regressor – team 'Text Mining'). The majority used deep learning approaches (e.g., BERT, RoBERTa) or other large pre-trained transformer models (e.g., teams "IITK", "Fired_from_NLP, HausaNLP"). When using these models, the teams would often experiment with different hyperparameters. Some teams went further and modified the specific learning approach or representations learned through methods such as contrastive learning (e.g., team: IITK).

5.4.3 Most Effective and Original Methods

In track A, the participants used the provided training sets for each of the 9 languages included in the track (`amh, arq, ary, eng, esp, hau, kin, mar` and `tel`). Overall, the different teams explored several approaches to enhance the performance. For instance, the top performing team PALI, used MT-DNN (Multi-Task Deep Neural Networks for Natural Language Understanding) (Liu et al., 2019a) and outperformed all the other teams across all languages except for Spanish and Kinyarwanda. For Kinyarwanda, king001 who used MT for data augmentation and multilingual mixed training and XLM-R (Conneau et al., 2020) as a base model achieved the best performance, and AAdaM who used translation-based data augmentation and adapter-based tuning reported the best score.

Note. however, that since PALI and king001 did not submit system description papers, they are not ranked in Tables 3 and 4.

5.5 Task B: Unsupervised

5.5.1 Best Performing Systems

SATLab Team SATLab used a system based on a model developed for authorship identification of source code (Bestgen, 2019). The system processed each pair of utterances independently, generating a distance between them without relying on additional information. Their pre-processing involved lower-casing of texts and making use of character n-grams ranging from 1 to 5 characters, encompassing all characters including spaces, punctuation marks, symbols, and characters from different writing systems. All n-grams were retained without a frequency threshold. The frequency of each feature was weighted by a logarithmic function, and the features of each statement were weighted by the L2 norm. The semantic similarity between utterances was estimated using the Euclidean distance between sets of n-grams in each pair.

MasonTigers In the initial phase, team MasonTigers obtained the embeddings of training data instances and used TF–IDF, PPMI, LaBSE sentence transformer, and language-specific BERT models for multiple languages. Cosine similarity scores were then computed between pairs of embeddings, followed by the use of ElasticNet and Linear Regression separately to predict sentence pair similarity. Predicted values were clipped to ensure a range from 0 to 1.

HW–TSC Team HW-TSC's method included the N-gram chars utilising tokenizers from XLM-RoBERTa and m-BERT as key features to compute similarity scores based on n-gram dictionaries of sentences. They also used BERTScore to assess text quality based on the cosine similarity of token-level representations from the BERT model.

5.5.2 Popular Methods

As the main challenge with track B was the prevention of using any data of more than two words long related to semantics, many teams such as Hausa–NLP and Tübingen–CL used pre-trained language models such as All-MiniLM-L6-v2 (Reimers and Gurevych, 2019).

Most teams opted for language-specific data and models, if not trained on similarity data, and compared the performance to monolingual BERT models. However, none of these methods were used by the top three performing teams.

5.5.3 Most effective and Original Methods

The most effective methods for the unsupervised track for all languages were submitted by teams SATLab, MasonTigers, and HW–TSC (top–3). SATLab's approach involved processing pairs independently using character n-grams. MasonTigers, on the other hand, leveraged various embedding methods and statistical machine learning using simple features such as TF-IDF and BERT models to compute the cosine similarity between embeddings, further refined using ElasticNet. On the other hand, The HW–TSC team used innovative techniques

Rank	Team	`afr`	`amh`	`arb`	`arq`	`ary`	`eng`	`esp`	`hau`	`hin`	`ind`	`kin`	`pun`	`Average`
1	SATLab (Bestgen, 2024)	0.761	**0.764**	0.487	**0.521**	**0.599**	0.774	**0.709**	**0.513**	0.649	**0.491**	0.458	-0.215	**0.543**
2	MasonTigers (Goswami et al., 2024)	0.757	0.656	0.405	0.424	0.561	0.766	0.661	0.504	0.571	0.382	**0.465**	0.020	0.514
3	HW–TSC (Piao et al., 2024)	0.639	0.650	0.402	0.296	0.460	0.758	0.641	0.382	0.613	0.445	0.323	**0.173**	0.482
4	UAlberta (Shi et al., 2024)	**0.789**	0.723	0.467	0.368	0.063	0.775	0.680	0.380	0.691	0.484	0.378	-0.027	0.481
*	Lexical Overlap	0.706	0.633	0.320	0.400	0.627	0.670	0.670	0.306	0.527	0.553	0.333	-0.274	0.456
5	silp_nlp (Singh et al., 2024)	0.732	0.643	0.314	0.402	0.552	0.317	-	0.387	0.571	0.532	0.350	-0.110	0.400
6	HausaNLP (Salahudeen et al., 2024)	0.716	0.038	0.202	0.334	0.397	0.819	0.618	0.358	0.440	0.407	0.404	-0.084	0.387
*	SemRel baseline (XLMR)	0.562	0.573	0.316	0.247	0.174	0.601	0.689	0.041	0.507	0.467	0.132	-0.072	0.353
NR	IITK (Basak et al., 2024)	-	0.068	-	0.489	0.358	0.808	0.591	0.379	-	-	-	-	0.449
NR	YSP (Aali et al., 2024)	-	-	-	0.385	-	0.788	0.598	0.193	-	-	0.377	-	0.468
NR	Tübingen-CL (Zhang and Çöltekin, 2024)	-	-	-	-	-	**0.837**	0.705	-	0.649	-	-	-	0.730
NR	CAILMD-23 (Sonavane et al., 2024)	-	-	-	-	-	0.819	-	-	**0.797**	-	-	-	0.808
NR	Self-StrAE (Opper and Narayanaswamy, 2024)	0.765	-	-	-	-	-	0.635	-	-	-	-	-	0.700
NR	NLU-STR (Malaysha et al., 2024)	-	-	**0.489**	-	-	-	-	-	-	-	-	-	0.489

Table 5: Track B results. The best results are in bold, and NR stands for *not ranked*. As the methods are highly language-dependent, we only rank teams that participated in at least 8 sub-tracks, but we highlight in blue the best results achieved by non-ranked teams. (Non-ranked teams are sorted based on the number of languages they participated in.)

such as the N-gram chars method with XLM-R and m-BERT tokenizers, as well as the BERTScore to evaluate the text quality.

In Table 5, we also have honorable mentions for teams that did not participate in all the languages but achieved remarkable results in one or a few languages. Notably, team CAILMD–23 achieved the best results in Hindi by using Hindi-BERT-v2, and team Tübingen–CL achieved the best results in English.

5.6 Task C: Crosslingual

5.6.1 Best Performing Systems

AAdaM They experimented with full fine-tuning, adapter fine-tuning using MAD (Pfeiffer et al., 2020), and data augmentation using different language combinations to augment data in a given source language.

UAlberta They used an XGBoost regressor-based (Chen and Guestrin, 2016) ensemble approach to integrate the predicted relatedness scores of three distinct regression models, with one optional SBERT model, as input and returned the final relatedness score as output. They applied the English version of their method trained for Track A to the translations of the non-English test sets. The regression model fine-tuned on MPNet was used in the XGBoost ensemble only for `amh`, `hau`, and `hin`, but not for the other languages such as `esp`, `ary`, `kin`, `ind`, `arb`, `arq`, and `afr`. The pre-trained English language models that were used include RoBERTa Large, T5 Base, and GPT2 Base, as well as MPNet only for languages `amh`, `hau`, and `hin`.

silp_nlp They used the provided datasets and cross-lingual transferability with all the provided datasets, except data in the target language, as a source. Their cross-lingual transfer approach made use of MuRIL (Khanuja et al., 2021) which led to the best results for Hindi and XLM-R (Conneau et al., 2020) led to the best ones for all the other languages.

5.6.2 Popular Methods

For the crosslingual track, many teams including best-performing ones such as UAlberta chose approaches similar to the ones used for supervised sub-tracks (e.g., using an XGBoost regressor (Chen and Guestrin, 2016)). As the main challenge was to determine how to leverage data in languages other than the target, many teams combined the provided SemRel datasets in all possible languages (e.g., king001, AAdaM). Some used the training datasets without any modifications (e.g., team HausaNLP) and others experimented with different language combinations to use those that would lead to the best results (e.g., MasonTigers). Finally, some teams applied advanced techniques to modify the vector embedding space (e.g., by adjusting for the anisotropic nature of vector spaces – team: USTC-CTSU).

5.6.3 Most Effective and and Original Methods

Overall, applying methods that are similar to the ones used in the supervised track using data in different languages can indeed lead to good results (e.g., king001, AAdaM, UAlberta). In addition, combining data in different languages and testing on another could boost the performance of crosslin-

Rank	Team	afr	amh	arb	arq	ary	eng	esp	hau	hin	ind	kin	pun	Average
1	AAdaM (Zhang et al., 2024)	0.814	0.863	0.653	0.551	0.600	0.794	0.621	0.729	0.839	**0.528**	0.650	**0.155**	**0.650**
2	UAlberta (Shi et al., 2024)	0.806	0.816	**0.671**	0.441	0.602	-	0.572	0.678	0.828	0.449	0.636	-0.017	0.589
*	SemRel baseline (LaBSE)	0.786	0.838	0.615	0.463	0.404	0.800	0.623	0.625	0.760	0.472	0.571	-0.049	0.579
3	silp_nlp (Singh et al., 2024)	0.747	0.805	0.427	0.387	0.673	0.737	0.569	0.643	0.801	0.472	-	-0.037	0.566
4	MaiNLP (Zhou et al., 2024)	0.738	0.728	0.399	0.274	0.568	-	-	-	0.695	0.319	**0.681**	0.087	0.499
5	USTCCTSU (Li et al., 2024a)	0.603	0.656	0.469	0.420	0.402	0.700	0.689	0.111	0.596	0.476	0.302	-0.084	0.445
6	umbclu (Roy Dipta and Vallurupalli, 2024)	**0.822**	0.043	0.035	0.126	-0.038	0.788	0.609	0.457	0.155	0.515	0.484	-0.078	0.326
7	HausaNLP (Salahudeen et al., 2024)	0.737	-0.031	0.184	0.074	0.276	0.360	0.604	0.177	0.346	0.472	0.319	0.114	0.303
8	MasonTigers (Goswami et al., 2024)	0.385	0.131	0.213	0.221	0.203	0.310	0.557	0.099	0.511	0.133	0.079	0.020	0.239
NR	USTC_NLP	0.749	0.709	0.517	0.414	0.613	0.784	0.685	0.476	0.658	0.460	0.454	-0.248	0.523
NR	king001	0.810	0.878	0.657	0.614	0.820	-	0.708	0.733	0.844	0.376	0.630	-0.050	0.641
NR	saturn	0.818	0.814	-	-	-	-	-	0.569	-	-	0.604	-0.103	0.540
NR	YSP (Aali et al., 2024)	-	-	-	0.225	-	0.819	0.657	0.212	-	-	0.256	-	0.434
NR	CAILMD-23 (Sonavane et al., 2024)	-	-	-	-	-	0.786	-	-	0.810	-	-	-	0.798
NR	PALI	-	-	-	-	0.842	-	-	-	-	-	-	-	0.842
NR	faridlazuarda	-	-	-	-	-	-	-	-	-	-	0.600	0.058	0.329
NR	ETMS@IITKGP	-	-	-	-	-	-	0.549	-	-	-	-	-	0.549
NR	Silp_nlp	-	-	-	-	-	-	-	-	-	0.472	-	-	0.472
NR	lukmanaj	-	-	-	-	-	-	-	0.177	-	-	-	-	0.177

Table 6: Track C results. The best results are in bold, and NR stands for *not ranked*. As the methods are highly language-dependent, we only rank teams that participated in at least 8 sub-tracks, but we highlight in blue the best results achieved by non-ranked teams. (Non-ranked teams are sorted based on the number of languages they participated in.)

gual models for STR as shown by team sil_nlp who achieved the best results in Amharic and Moroccan Arabic. Further, we note that leveraging advanced features such as (1) linguistic features (e.g., language family) as performed by MaiNLP, who achieved the best results for Kinyarwanda, and (2) embedding features by adjusting the distribution of the similarity scores as experimented by USTC-CTSU could also help boost the performance.

Besides reporting on the best-performing teams only, in Table 6, we also mention teams that did not participate in many sub-tracks but achieved good results such as team YSP, which outperforms all the other teams in English.

6 Discussion

We observe that in general, teams opt out of pre-trained models, and in most cases, the methods do not perform equally well across languages. Hence, for a given track, performing well in a language does not mean performing equally well in another language.

Further, the results show that good scores are not only related to low vs. high-resourcedness. For instance, In tracks B and C, results for Modern Standard Arabic (arb), which is considered high resource, are sometimes worse than those for low resource languages such as Amharic (amh) and Kinyarwanda (kin).

Interestingly, although the participating teams rarely use language-specific features, such approaches lead to good and interpretable results, as reported by e.g., team MaiNLP, who leveraged information about language families in Track C. We also note that for Track C, using a simple LaBSE baseline can achieve results that are better or comparable to more sophisticated techniques (see Ousidhoum et al. (2024) for language-specific baseline results).

7 Conclusion

We presented the first shared task on semantic relatedness, covering three tracks and 14 languages in total. The submitted systems were ranked based on the ranking of their predicted relatedness scores compared to the gold labels.

We summarised the reported results, the best-performing methods, and the most effective, promising, and original ones. Overall, our findings on sentence representation techniques vary across the different languages and show that determining semantic textual relatedness is not a trivial task.

8 Limitations

As stated in Ousidhoum et al. (2024), we acknowledge that there is no formal definition of what constitutes semantic relatedness and that our 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 it, 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. However, they remain a good starting point as they were carefully picked, labeled, and processed by native speakers.

9 Ethics Statement

As stated in Ousidhoum et al. (2024), 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 as these are not fully representative of its usage.

References

Yasamin Aali, Sardar Hamidian, and Parsa Farinneya. 2024. Ysp at semeval-2024 task 1: Enhancing sentence relatedness assessment using siamese networks. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 946–950, Mexico City, Mexico. Association for Computational Linguistics.

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.

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.

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.

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. 2013. *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.

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.

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.

Senthil Kumar B, Aravindan Chandrabose, Gokulakrishnan B, and Karthikraja TP. 2024. NLP_Team1@SSN at semeval-2024 task 1: Impact of language models in sentence-bert for semantic textual relatedness in low-resource languages. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1866–1871, Mexico City, Mexico. Association for Computational Linguistics.

Udvas Basak, Rajarshi Dutta, Shivam Pandey, and Ashutosh Modi. 2024. IITK at semeval-2024 task 1: Contrastive learning and autoencoders for semantic textual relatedness in multilingual texts. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1454–1459, Mexico City, Mexico. Association for Computational Linguistics.

Abdessamad Benlahbib, Anass Fahfouh, Hamza Alami, and Achraf Boumhidi. 2024. NLP-LISAC at semeval-2024 task 1: Transformer-based approaches for determining semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 213–217, Mexico City, Mexico. Association for Computational Linguistics.

Yves Bestgen. 2019. CECL at SemEval-2019 task 3: Using surface learning for detecting emotion in textual conversations. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 148–152, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yves Bestgen. 2024. Satlab at semeval-2024 task 1: A fully instance-specific approach for semantic textual relatedness prediction. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 95–100, Mexico City, Mexico. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. 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.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM.

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.

Fahad Ebrahim and Mike Joy. 2024. WarwickNLP at semeval-2024 task 1: Low-rank cross-encoders for efficient semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 246–252, Mexico City, Mexico. Association for Computational Linguistics.

Seyedeh Fatemeh Ebrahimi, Karim Akhavan Azari, Amirmasoud Iravani, Hadi Alizadeh, Zeinab Taghavi, and Hossein Sameti. 2024. Sharif-STR at semeval-2024 task 1: Transformer as a regression model for fine-grained scoring of textual semantic relations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1032–1041, Mexico City, Mexico. Association for Computational Linguistics.

Alex Eponon and Luis Ramos Perez. 2024. Pinealai at semeval-2024 task 1: Exploring semantic relatedness prediction using syntactic, tf-idf, and distance-based features. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 922–926, Mexico City, Mexico. Association for Computational Linguistics.

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

T. N. Flynn and A. A. J. Marley. 2014. Best-worst scaling: theory and methods. In Stephane Hess and Andrew Daly, editors, *Handbook of Choice Modelling*, pages 178–201. Edward Elgar Publishing.

Shreejith G, Ravindran V, Aashika Jetti, Rajalakshmi Sivanaiah, and Angel Deborah S. 2024. Techssn at semeval-2024 task 1: Multilingual analysis for semantic textual relatedness using boosted transformer models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 894–899, Mexico City, Mexico. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dhiman Goswami, Sadiya Sayara Chowdhury Puspo, Md Nishat Raihan, Al Nahian Bin Emran, Amrita Ganguly, and Marcos Zampieri. 2024. Masontigers at semeval-2024 task 1: An ensemble approach for semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1370–1380, Mexico City, Mexico. Association for Computational Linguistics.

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

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Md. Sajjad Hossain, Ashraful Islam Paran, Symom Hossain Shohan, Jawad Hossain, and Mohammed Moshiul Hoque. 2024. SemanticCUET-Sync at semeval-2024 task 1: Finetuning sentence transformer to find semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1212–1218, Mexico City, Mexico. Association for Computational Linguistics.

Tollef Jørgensen. 2024. PEAR at semeval-2024 task 1: Pair encoding with augmented re-sampling for semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1395–1401, Mexico City, Mexico. Association for Computational Linguistics.

Ron Keinan. 2024. Text mining at semeval-2024 task 1: Evaluating semantic textual relatedness in low-resource languages using various embedding methods and machine learning regression models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 407–418, Mexico City, Mexico. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha P. Talukdar. 2021. Muril: Multilingual representations for indian languages. *CoRR*, abs/2103.10730.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, San Diego, California.

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

Svetlana Kiritchenko and Saif M. Mohammad. 2017b. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of The Annual Meeting of the Association for Computational Linguistics (ACL)*, Vancouver, Canada.

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

Jianjian Li, Shengwei Liang, Yong Liao, Hongping Deng, and Haiyang Yu. 2024a. USTCCTSU at semeval-2024 task 1: Reducing anisotropy for cross-lingual semantic textual relatedness task. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 868–874, Mexico City, Mexico. Association for Computational Linguistics.

Weijie Li, Jin Wang, and Xuejie Zhang. 2024b. Ynu-hpcc at semeval-2024 task 1: Self-instruction learning with black-box optimization for semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 779–786, Mexico City, Mexico. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach.

Jordan J Louviere and George G Woodworth. 1991. Best-Worst Scaling: A Model For The Largest Difference Judgments. Technical report, Working paper.

Anand Kumar M and Hemanth Kumar M. 2024. scalar semeval-2024 task 1: Semantic textual relatednes for english. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 889–893, Mexico City, Mexico. Association for Computational Linguistics.

Sanad Malaysha, Mustafa Jarrar, and Mohammed Khalilia. 2024. NLU-STR at semeval-2024 task 1: Generative-based augmentation and encoder-based scoring for semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 881–888, Mexico City, Mexico. Association for Computational Linguistics.

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.

Daniela Moctezuma, Eric Tellez, and Mario Graff. 2024. Ingeotec at semeval-2024 task 1: Bag of words and transformers. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1144–1148, Mexico City, Mexico. Association for Computational Linguistics.

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

Anderson Morillo, Daniel Peña, Juan Carlos Martinez Santos, and Edwin Puertas. 2024. Verbanexai lab at semeval-2024 task 1: A multilayer artificial intelligence model for semantic relationship detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1334–1340, Mexico City, Mexico. Association for Computational Linguistics.

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.

Kiet Nguyen and Dang Thin. 2024. NRK at semeval-2024 task 1: Semantic textual relatedness through domain adaptation and ensemble learning on bert-based models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 76–81, Mexico City, Mexico. Association for Computational Linguistics.

Mattia Opper and Siddharth Narayanaswamy. 2024. Self-StrAE at semeval-2024 task 1: Making self-structuring autoencoders learn more with less. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 108–115, Mexico City, Mexico. Association for Computational Linguistics.

Bryan Orme. 2009. Maxdiff analysis: Simple counting, individual-level logit, and HB. Sawtooth Software, Inc.

Jesus-German Ortiz-Barajas, Gemma Bel-Enguix, and Helena Goméz-Adorno. 2024. MBZUAI-UNAM at semeval-2024 task 1: Sentence-crobi, a simple cross-bi-encoder-based neural network architecture for semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1060–1068, Mexico City, Mexico. Association for Computational Linguistics.

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, and Saif M. Mohammad. 2024. Semrel2024: A collection of semantic textual relatedness datasets for 14 languages.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Mengyao Piao, Su Chang, Yuang Li, Xiaosong Qiao, Xiaofeng Zhao, Yinglu Li, Min Zhang, and Hao Yang. 2024. Hw-tsc 2024 submission for the semeval-2024 task 1: Semantic textual relatedness (str). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1645–1649, Mexico City, Mexico. 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.

Shubhashis Roy Dipta and Sai Vallurupalli. 2024. UMBCLU at semeval-2024 task 1: Semantic textual relatedness with and without machine translation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1341–1347, Mexico City, Mexico. Association for Computational Linguistics.

Saheed Abdullahi Salahudeen, Falalu Ibrahim Lawan, Yusuf Aliyu, Amina Abubakar, Lukman Aliyu, Nur Bala Rabiu, Mahmoud Said Ahmad, Idi Mohammed, Aliyu Rabiu Shuaibu, Alamin Musa, Auwal Shehu Ali, and Zedong Nie. 2024. Hausanlp at semeval-2024 task 1: Textual relatedness analysis for semantic representation of sentences. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 188–192, Mexico City, Mexico. Association for Computational Linguistics.

Anik Shanto, Md. Sajid Alam Chowdhury, Mostak Chowdhury, Udoy Das, and Hasan Murad. 2024. Fired_from_NLP at semeval-2024 task 1: Towards developing semantic textual relatedness predictor: A transformer-based approach. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 846–851, Mexico City, Mexico. Association for Computational Linguistics.

Ning Shi, Senyu Li, Guoqing Luo, Amirreza Mirzaei, Ali Rafiei, Jai Riley, Hadi Sheikhi, Mahvash Siavashpour, Mohammad Tavakoli, Bradley Hauer, and Grzegorz Kondrak. 2024. UAlberta at semeval-2024 task 1: A potpourri of methods for quantifying multilingual semantic textual relatedness and similarity. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1810–1817, Mexico City, Mexico. Association for Computational Linguistics.

Marco Siino. 2024. All-mpnet at semeval-2024 task 1: Application of mpnet for evaluating semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 372–377, Mexico City, Mexico. Association for Computational Linguistics.

Sumit Singh, Pankaj Kumar Goyal, and Uma Shanker Tiwary. 2024. silp_nlp at semeval-2024 task 1: Cross-lingual knowledge transfer for mono-lingual learning. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1187–1193, Mexico City, Mexico. Association for Computational Linguistics.

Srushti Sonavane, Sharvi Endait, Ridhima Sinare, Pritika Rohera, Advait Naik, and Dipali Kadam. 2024. Cailmd-23 at semeval-2024 task 1: Multilingual evaluation of semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 969–974, Mexico City, Mexico. Association for Computational Linguistics.

Lianshuang Su and Xiaobing Zhou. 2024. Nlp_str_teams at semeval-2024 task1: Semantic textual relatedness based on mask prediction and bert model. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 330–334, Mexico City, Mexico. Association for Computational Linguistics.

Hidetsune Takahashi, Xingru Lu, Sean Ishijima, Deokgyu Seo, Yongju Kim, Sehoon Park, Min Song, Kathylene Marante, Keitaro-Luke Iso, Hirotaka Tokura, and Emily Ohman. 2024. OZemi at semeval-2024 task 1: A simplistic approach to textual relatedness evaluation using transformers and machine translation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 7–12, Mexico City, Mexico. Association for Computational Linguistics.

Dilip Venkatesh and Sundaresan Raman. 2024. Bits pilani at semeval-2024 task 1: Using text-embedding-3-large and labse embeddings for semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 852–855, Mexico City, Mexico. Association for Computational Linguistics.

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.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leixin Zhang and Çağrı Çöltekin. 2024. Tübingen-CL at semeval-2024 task 1: Ensemble learning for semantic relatedness estimation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1008–1014, Mexico City, Mexico. Association for Computational Linguistics.

Miaoran Zhang, Mingyang Wang, Jesujoba Alabi, and Dietrich Klakow. 2024. Aadam at semeval-2024 task 1: Augmentation and adaptation for multilingual semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 787–797, Mexico City, Mexico. Association for Computational Linguistics.

Shijia Zhou, Huangyan Shan, Barbara Plank, and Robert Litschko. 2024. MaiNLP at semeval-2024 task 1: Analyzing source language selection in cross-lingual textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1854–1865, Mexico City, Mexico. Association for Computational Linguistics.

A Appendix: Track A–Best Performing Teams

PALI and king001 Both teams PALI and king001 did not submit a task description paper. king001 chose to use translation for data augmentation and multilingual mixed training. The team used XLM–R as their base model and DeBERTa–v3 (He et al., 2021).

AAdaM Team AAdaM opted for translation-based data augmentation to increase the training data size for better performance. The English STR training data and STSB (semantic similarity) data were translated to create augmented datasets in other languages. The team explored both fine-tuning and adapter-based tuning, aiming to examine and compare their effectiveness on STR across the different languages. Given a target language, they first fine-tuned the cross-encoder-based AfroXLMR model on the augmented data as a warm-up or TAPT (Task-Adaptive-Pre-Training) and then continued the fine-tuning on the provided STR data.

NRK They used ensembling and various BERT-like models.

PEAR They examined the effect of combining vs. using language-specific data through 5-fold validation. No text preprocessing was conducted to maintain fairness across languages. Three model configurations were defined: “base” with no training, “all” trained on all languages, and “lang” trained on one language. They experimented with multilingual embeddings, cross-encoders, and data augmentation with bi-encoders. Parameter optimization was conducted using Optuna.

silp_nlp Team silp_nlp’s methodology for track A was a two-stage training. In the initial stage, they trained a model using all 9 languages covered in track A with MuRIL (Khanuja et al., 2021). They experimented with different hyperparameters on five epochs and selected the best multilingual checkpoint based on the average validation data loss. They fine-tuned the resulting model using the training data for each language in track A and ended up with monolingual models.

Each monolingual model was trained using different hyperparameters and they selected their final model based on the validation data loss of the corresponding language track.

NLP_1@SSN They used SBERT fine-tuned on multilingual and monolingual pre-trained language models Overall, they observed that the usage of monolingual PLMs did not guarantee better performance.

UAlberta They used an ensemble approach with an XGBoost regression (Chen and Guestrin, 2016) to integrate the predicted relatedness scores of three distinct regression models, with one optional SBERT model, as input and returned the final relatedness scores as output. Each of these models used a different pre-trained language model as its backbone, specifically RoBERTa Large (Liu et al., 2019b), T5 Base, GPT-2 Base, and the optional SBERT (MPNet). They merged the English training and development sets with the translated training set of the target language. Then, they split them again via uniform random sampling according to their original sizes to establish new training and development splits. The did not use the data provided for arq, ary, and kin, and applied the English-trained version of their method to the English translations of the arq, ary, and kin test sets instead.

MBZUAI-UNAM They fine-tuned a paraphrase model architecture to train language-specific models, using a separate pre-trained model to embed each language. They also experimented with combined training sets based on the language families.

INGEOTEC For English and Spanish, they used embeddings (microsoft/mpnet-base, bert-base-multilingual-cased) to train an SVM classifier. For the other languages, they used prior work EvoMSA.

HausaNLP They used different base pre-trained models.

B Appendix: Track B

SATLab They proposed a system based on a model developed for the authorship identification of source code (Bestgen, 2019). It processed each pair of utterances independently, generating a distance between them without relying on additional information. Pre-processing involved lower-casing of texts. Character n-grams ranging from 1 to 5 characters are used, encompassing all characters including spaces, punctuation marks, symbols, and characters from different writing systems, all n-grams are retained without a frequency threshold. The frequency of each feature was weighted by a logarithmic function, and the features of each statement were weighted by the L2 norm. Semantic similarity between utterances was estimated using Euclidean distance between sets of n-grams in each pair.

MasonTigers In the initial phase, team MasonTigers obtained embeddings of training data and used various methods including TF-IDF, PPMI, LaBSE sentence transformer, and language-specific BERT models for multiple languages. Cosine similarity was then computed between pairs of embeddings, followed by applying ElasticNet and Linear Regression separately to predict sentence pair similarity in the development phase. Predicted values were clipped to ensure a range from 0 to 1.

HW–TSC The key features used by team HW-TSC's method included the N-gram chars method using XLM-RoBERTa and m-BERT tokenizers to compute similarity scores based on n-gram sentence dictionaries. They also used the BERTScore method to assess text quality based on the cosine similarity of token-level representations from the BERT model.

UAlberta They used a linear combination of two sets of normalized results, each derived from the cosine similarity measurements of sentence embeddings obtained from the hidden sentence representations processed by BERT Large and RoBERTa Large. They calculated the final relatedness scores by averaging the cosine similarity scores of sentence embeddings obtained from each set.

silp_nlp They converted the sentences into unigram and bigram representations and used Support Vector Regression (SVR).

Sentences were combined and transformed into a vector, and each sentence was indexed based on a value that represented the count of unigrams/bigrams present in it. The resulting vector was fed into the SVR model along with label values for training.

HausaNLP Team HausaNLP used a standard all-MiniLM-L6-v2 model to train a model for Track B.

IITK Team IITK uses SimCSE (Gao et al., 2021), or Simple Contrastive Learning of Sentence Embeddings that induced slight variations in its representation through dropout. TSDAE(Wang et al., 2021), a denoising autoencoder, was used to generate sentence embeddings by reconstructing original sentences in the presence of noise. They used BERT to construct the denoising autoencoder and TSDAE optimized the likelihood of reconstructing sentences during training, which led to compact embeddings.

Tübingen-CL Team Tübingen-CL opted for exploring features like cosine distance of average word embeddings and word overlap ratios, to potentially enhance performance. For English, they used two models: multi-qa-MiniLM-L6-cos-v1 trained on QA pairs and trained for semantic search and e5-

base-unsupervised trained on various pairs including question-answer and post-comment pairs, both refined with unsupervised transformation (PCA). Two additional features, PCA-transformed GloVe embeddings, and content word overlap ratios were incorporated into the unsupervised ensemble system. Similar methods were applied for Spanish and Hind using multilingual BERT embeddings and various feature combinations to predict relatedness.

CAILMD-23 Team CAILMD-23 participated in the English and Hindi sub-tracks of the unsupervised task. They experimented with a few models such as BERT-based and Hindi-Bert v2. The latter is trained on Hindi text comprehension with a training corpus of roughly 1.8 billion tokens.

C Appendix: Track C

AAdaM hey experimented with full fine-tuning, adapter fine-tuning using MAD (Pfeiffer et al., 2020), and data augmentation using different language combinations to augment data in a given source language.

king001 They did not submit a system description paper but they reported combining the training datasets provided for track A, and if one of them was in the target language, they translated it into English. Then, they run multi-task learning for 15 epochs.

UAlberta They used an ensemble approach with an XGBoost regressor (Chen and Guestrin, 2016) to integrate the predicted relatedness scores of three distinct regression models, with one optional SBERT model, as input. Each of their models used a different pre-trained language model as its backbone, specifically RoBERTa Large, T5 Base, GPT-2 Base, and the optional SBERT (MPNet).

They applied the English version of their method reported for Track A to the translations of the non-English test sets. The regression model fine-tuned on MPNet was used in the XGBoost ensembling method for amh, hau, and hin and not for esp, ary, kin, ind, arb, arq, and afr.

silp_nlp They used cross-lingual transferability on all the provided datasets except for the target language (e.g., when they test on Telugu, they use all languages except Telugu). In their cross-lingual transfer approach, MuRIL (Khanuja et al., 2021) led to the best results for Hindi and XLM-R (Conneau et al., 2020) for all the other languages.

USTCCTSU They used XLM-R (Conneau et al., 2020) trained on a combination of language inputs (chosen by trying different combinations with the best one including all the languages). They ranked in the top 5 for ind,arq, and esp.

They adjusted the similarity scores for the XLM-R base models by applying a technique called *whitening* that allowed them to change the non-uniform score distribution into multiple distributions, and eventually, into a uniform one.

MaiNLP They finetuned multilingual LLMs (XLM-R and Furina) using an upscaled version of the data from Track A. They assessed the linguistic similarity of the available Track A data to determine the most useful datasets and experimented with different language families. For pre-processing, they used tokenization, segmentation, and translation. They also experimented with transliteration to change the scripts into Latin. Translations helped them upscale the English, Hausa, and Spanish training data and then evaluate on the Track C data. They achieved the best results for Kinyarwanda.

umbclu They pre-trained T5 models with SemRel data. They used the English fine-tuned models for inference on all language test sets except English. On the other hand, they used Spanish models for inference on English.

HausaNLP They used a BERT-based model fine-tuned on the datasets in other languages. E.g., they trained on English data and tested on Spanish, trained on Kinyarwanda and tested on Hausa. They ranked in the top 5 in Task C for ind, pan.

MasonTigers They used statistical machine learning (Linear Regression, ElasticNet with TF-IDF and PPMI features) along with language-specific BERT-based models to predict the relatedness scores. The models were trained on dataset combinations of 5 languages other than the target language and used BERT-based models's similarity prediction on the target test data (e.g., they trained on amh, eng, esp, arq, ary and tested on afr). For language-specific BERT-like models, they used African language BERT-base models, Arabic BERT-based models, African-BERTa, and for eng, hin, ind, pun, esp, they used spanBERTa, BanglaBERT, RoBERTa-tagalog-base-BERT, HindiBERT, and RoBERTa.

Team	Paper
AAdaM	Zhang et al. (2024)
All–Mpnet	Siino (2024)
BITS Pilani	Venkatesh and Raman (2024)
CAILMD–23	Sonavane et al. (2024)
Fired_from_NLP	Shanto et al. (2024)
HausaNLP	Salahudeen et al. (2024)
HW-TSC	Piao et al. (2024)
IITK	Basak et al. (2024)
INGEOTEC	Moctezuma et al. (2024)
MaiNLP	Zhou et al. (2024)
MasonTigers	Goswami et al. (2024)
MBZUAI–UNAM	Ortiz-Barajas et al. (2024)
NLP–LISAC	Benlahbib et al. (2024)
NLP_STR_teamS	Su and Zhou (2024)
NLP_Team1SSN	B et al. (2024)
NLU–STR	Malaysha et al. (2024)
NRK	Nguyen and Thin (2024)
OZemi	Takahashi et al. (2024)
PEAR	Jørgensen (2024)
Pinealai	Eponon and Ramos Perez (2024)
SATLab	Bestgen (2024)
scaLAR	M and M (2024)
Self–StrAE	Opper and Narayanaswamy (2024)
SemanticCUETSync	Hossain et al. (2024)
Sharif_STR	Ebrahimi et al. (2024)
silp_nlp	Singh et al. (2024)
TECHSSN	G et al. (2024)
Text Mining	Keinan (2024)
Tübingen–CL	Zhang and Çöltekin (2024)
UAlberta	Shi et al. (2024)
UMBCLU	Roy Dipta and Vallurupalli (2024)
USTCCTSU	Li et al. (2024a)
VerbaNexAI	Morillo et al. (2024)
WarwickNLP	Ebrahim and Joy (2024)
YNU–HPCC	Li et al. (2024b)
YSP	Aali et al. (2024)

Table 7: The participating teams (alphabetically ordered) that submitted system description papers.