MUSTS: A Multilingual Semantic Textual Similarity Benchmark for Diverse Languages

Hansi Hettiarachchi and Tharindu Ranasinghe
School of Computing and Communications, Lancaster University, UK
{h.hettiarachchi, t.ranasinghe}@lancaster.ac.uk

Relevant UniDive working groups: WG3, WG4

1 Introduction

Semantic textual similarity (STS) measures the degree to which two sentences convey the same meaning (Cer et al., 2017) and underpins numerous NLP applications, including machine translation, summarisation, question answering, and information retrieval. While various benchmarks exist to evaluate STS methods, they suffer from critical limitations: limited coverage of low-resource languages, conflation of similarity with relatedness in annotation, and reliance on machine-translated data without post-editing. In particular, the widely adopted Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023) contains no low-resource languages in its STS component, leaving it unclear how top-ranked models generalise to underrepresented languages.

To address these gaps, we introduce MUSTS (MUltilingual Semantic Textual Similarity Benchmark), spanning 13 languages across 7 language families, including low-resource languages such as Sinhala, Tamil, and Serbian. This work, accepted at ACL 2025, contributes to UniDive's objectives of quantifying linguistic diversity in NLP resources (WG4) and advancing multilingual language technology (WG3).

2 The MUSTS Benchmark

Unlike existing benchmarks, MUSTS was compiled with three strict criteria: (i) all datasets follow established STS annotation guidelines (Agirre et al., 2013), excluding those annotated for relatedness; (ii) machine-translated datasets are included only if they underwent human post-editing; and (iii) all datasets were published in peer-reviewed venues. Table 1 summarises the benchmark.

MUSTS covers 3 high-resource, 7 mid-resource, and 3 low-resource languages following the taxonomy of Joshi et al. (2020), substantially improving upon MTEB's language coverage for STS evaluation.

Language	Family	|Train|	|Test|
Arabic	Semitic	1,080	250
Br. Portuguese	Romance	2,500	2,000
Czech	Slavic	925	500
English	Germanic	5,750	1,380
French	Romance	600	410
Japanese	Japonic	12,500	1,460
Korean	Koreanic	5,750	1,380
Portuguese	Romance	2,500	2,000
Romanian	Romance	5,750	1,380
Serbian	Slavic	953	239
Sinhala	Indo-Aryan	5,000	100
Spanish	Romance	1,620	250
Tamil	Dravidian	2,500	100

Table 1: MUSTS overview: 13 languages across 7 families, including low-resource languages (Serbian, Sinhala, Tamil).

3 Evaluation and Key Findings

We evaluated more than 25 models on MUSTS, spanning unsupervised methods (vector averaging, smooth inverse frequency, sentence embeddings, LLM prompting, LLM-based encoders) and supervised methods (fine-tuned transformers, trained LLM-based encoders), as summarised in Table 2. Four prompting techniques were used with LLMs: Zero-shot (ZS), Few-shot-English (FS-En), Few-shot-Monolingual (FS-Mono), and Few-shot with Chain of thoughts (FS-CoT). Models were assessed using Spearman correlation (r) between model predictions and human annotations across all 13 languages. Our key findings, relevant to UniDive's diversity objectives, are:

LLMs struggle with low-resource languages. LLM-based encoders and prompting approaches provide excellent results for mid- and high-resource languages but perform poorly on low-resource languages, particularly Sinhala and Tamil. For example, the top-ranked LLM-based encoder on MTEB, `NV-Embed-v2`, achieves strong average Spearman correlations across mid- and high-resource languages (e.g. $r = 0.843$ for English, $r = 0.857$ for Czech), yet drops substantially on Sinhala ($r = 0.343$) and Tamil ($r = 0.397$). By contrast, simpler methods outperform LLMs

Strategy	Model
Vector Averaging	XLM-R Large RemBERT infoXLM Large
Smooth Inverse Frequency	XLM-R Large RemBERT infoXLM Large
Sentence Embeddings	LaBSE MiniLM-L12-v2 LASER
LLM Prompting	Llama-3.1-8B-Instruct Mistral-8B-Instruct
Top LLM-based Encoders on MTEB	NV-Embed-v2 bge-en-icl stella_en_1.5B_v5 SFR-Embedding-2_R gte-Qwen2-7B-instruct gte-Qwen2-1.5B-instruct
Fine-tuned Transformers	XLM-R Large RemBERT infoXLM Large
Trained LLM-based Encoders	stella_en_1.5B_v5 gte-Qwen2-1.5B-instruct

Table 2: Models evaluated on MUSTS

for these languages: LaBSE (Feng et al., 2022), an earlier multilingual sentence encoder, achieves $r = 0.499$ on Sinhala and $r = 0.551$ on Tamil, and Smooth Inverse Frequency with infoXLM Large reaches $r = 0.426$ on Tamil, being competitive with or exceeding several LLM-based approaches. These results reveal that state-of-the-art models do not generalise equitably across the linguistic diversity spectrum.

MTEB rankings do not transfer to MUSTS. The model ranking on the MTEB leaderboard does not hold when evaluated on MUSTS. Among the top six LLM-based encoders from the MTEB leaderboard, SFR-Embedding-2_R, ranked fourth on MTEB, achieves the highest average Spearman correlation among unsupervised embedding models across all 13 languages ($r = 0.750$), outperforming the first-ranked MTEB model, NV-Embed-v2 ($r = 0.743$). The discrepancy is even more apparent for lower-ranked models: stella_en_1.5B_v5 and bge-en-icl lag considerably behind ($r = 0.663$ and $r = 0.668$, respectively), despite their higher positions in MTEB ranking. This discrepancy is primarily attributable to MUSTS's broader language coverage, reinforcing UniDive's advocacy for including low-resource languages in NLP evaluation.

Supervised training remains most effective. Fine-tuned transformer models yield the best overall STS performance across all 13 languages. Among these, infoXLM Large achieves the highest average Spearman correlation ($r = 0.880$), followed closely by RemBERT ($r = 0.852$) and XLM-R Large ($r = 0.846$), substantially outperforming all unsupervised approaches. Notably, supervised models also demonstrate strong performance on low-resource languages, with infoXLM Large reaching $r = 0.869$ on Serbian, $r = 0.825$ on Sinhala, and $r = 0.848$ on Tamil. Fine-tuning LLM-based encoders with sentence transformer architectures yields more modest but targeted gains. For instance, gte-Qwen2-1.5B-instruct improves from $r = 0.312$ (unsupervised) to $r = 0.420$ on Sinhala and from $r = 0.353$ to $r = 0.770$ on Tamil, suggesting that targeted adaptation strategies are needed for underrepresented languages.

4 Relevance to UniDive

MUSTS directly contribute to UniDive's mission along two axes. For **WG4 (Quantifying and promoting diversity)**, our benchmark provides a principled framework for measuring how STS performance varies across languages of different resource levels, empirically demonstrating the inter-linguistic diversity gap in a fundamental NLU task. The finding that MTEB rankings fail to generalise to MUSTS underscores the importance of diversity-aware evaluation, a core concern of UniDive. For **WG3 (Multilingual and cross-lingual language technology)**, MUSTS establishes the first dedicated multilingual STS benchmark with strict quality controls, enabling fairer comparison of cross-lingual NLP tools. The inclusion of Sinhala and Tamil, languages central to UniDive's low-resource language agenda, directly supports the Action's goal of better NLP coverage for underrepresented languages.

MUSTS is publicly available with evaluation code and a public leaderboard to facilitate ongoing multilingual STS research[1].

References

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

[1] https://github.com/TharinduDR/MUSTS

Conference and the Shared Task: Semantic Textual Similarity, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Daniel Cer, Mona Diab, 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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.