MTEB

Massive Text Embedding Benchmark

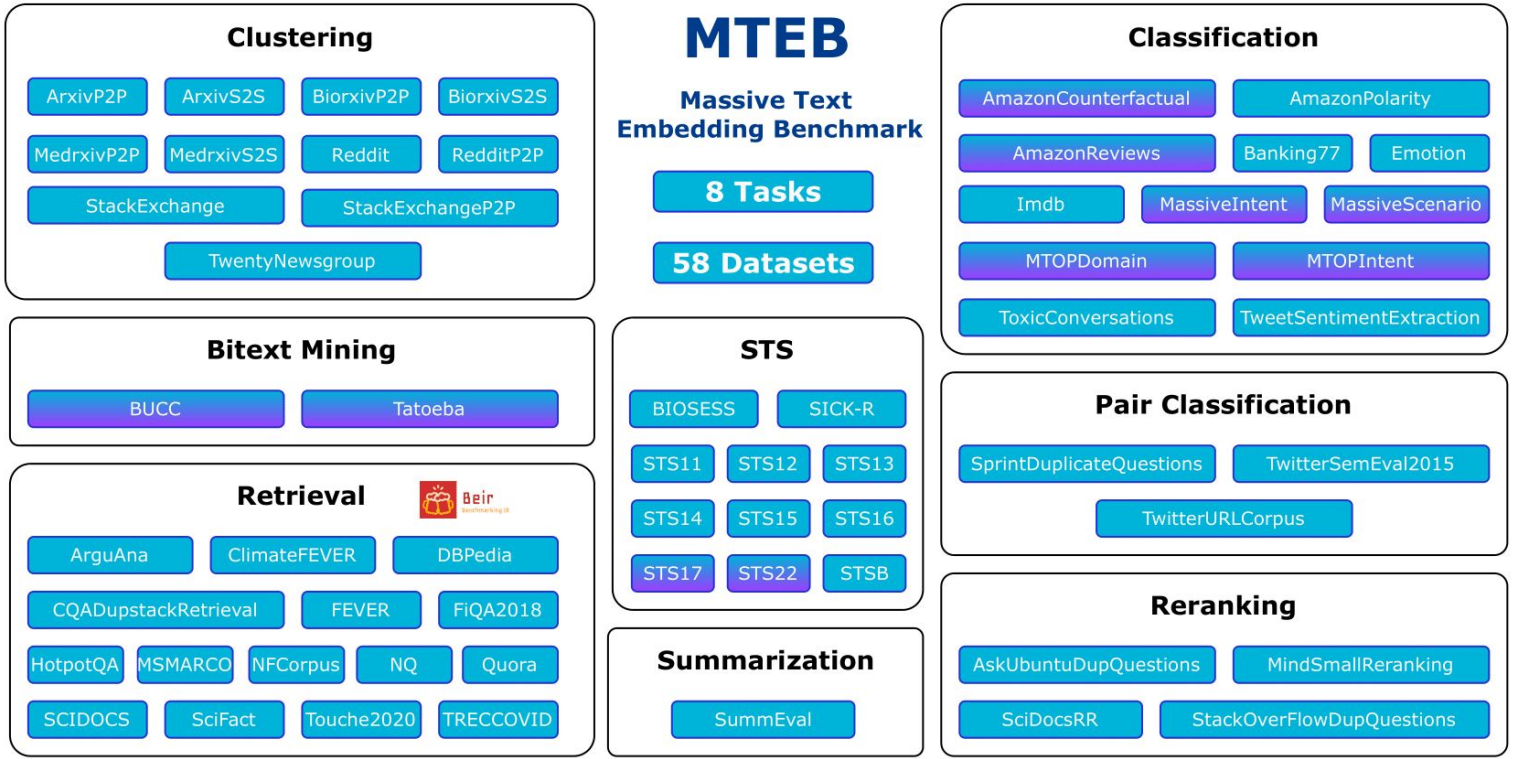
Niklas Muennighoff, Nouamane Tazi, Loïc Magne, Nils Reimers

Paper: arxiv.org/abs/2210.07316
Code: github.com/embeddings-benchmark/mteb
Data: hf.co/mteb
Leaderboard: hf.co/spaces/mteb/leaderboard

Hugging Face

co:here

I. DATA AND METRICS



Metrics

nDCG
Accuracy
Recall
Precision
AP
MAP
Precision
Recall
F1
MRR
V-Measure
Cosine Similarity
Dot Product
Euclidean Distance
Manhattan Distance
Spearman Correlation
Pearson Correlation

II. DESIDERATA

1. Diversity

2. Simplicity

- Benchmark in 5 lines of code

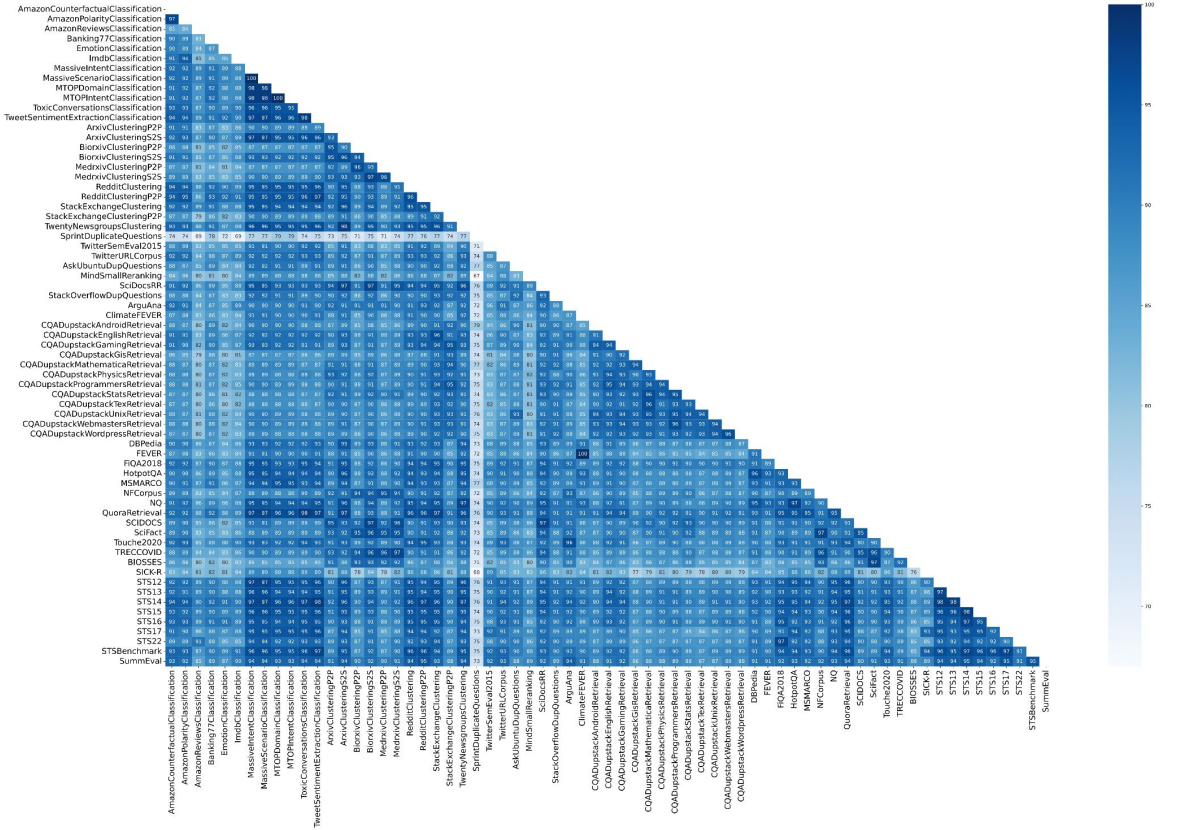
```python
from mteb import MTEB
from sentence_transformers import SentenceTransformer

# Define the sentence-transformers model name
model_name = "average_word_embeddings_komninos"

model = SentenceTransformer(model_name)
evaluation = MTEB(tasks=["Banking77Classification"])
results = evaluation.run(model, output_folder=f"results/{model_name}")
```

3. Extensibility

- 100% open-source
- Custom tasks / models

4. Reproducibility

- Versioning at dataset & library level



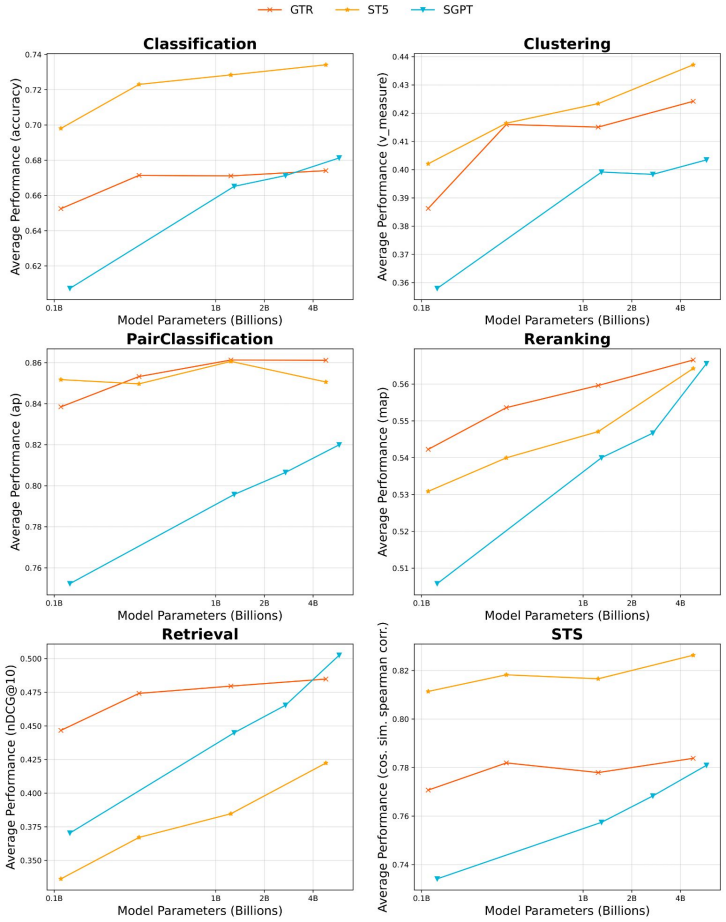
https://github.com/embeddings-benchmark/mteb

III. LARGE-SCALE BENCHMARKING

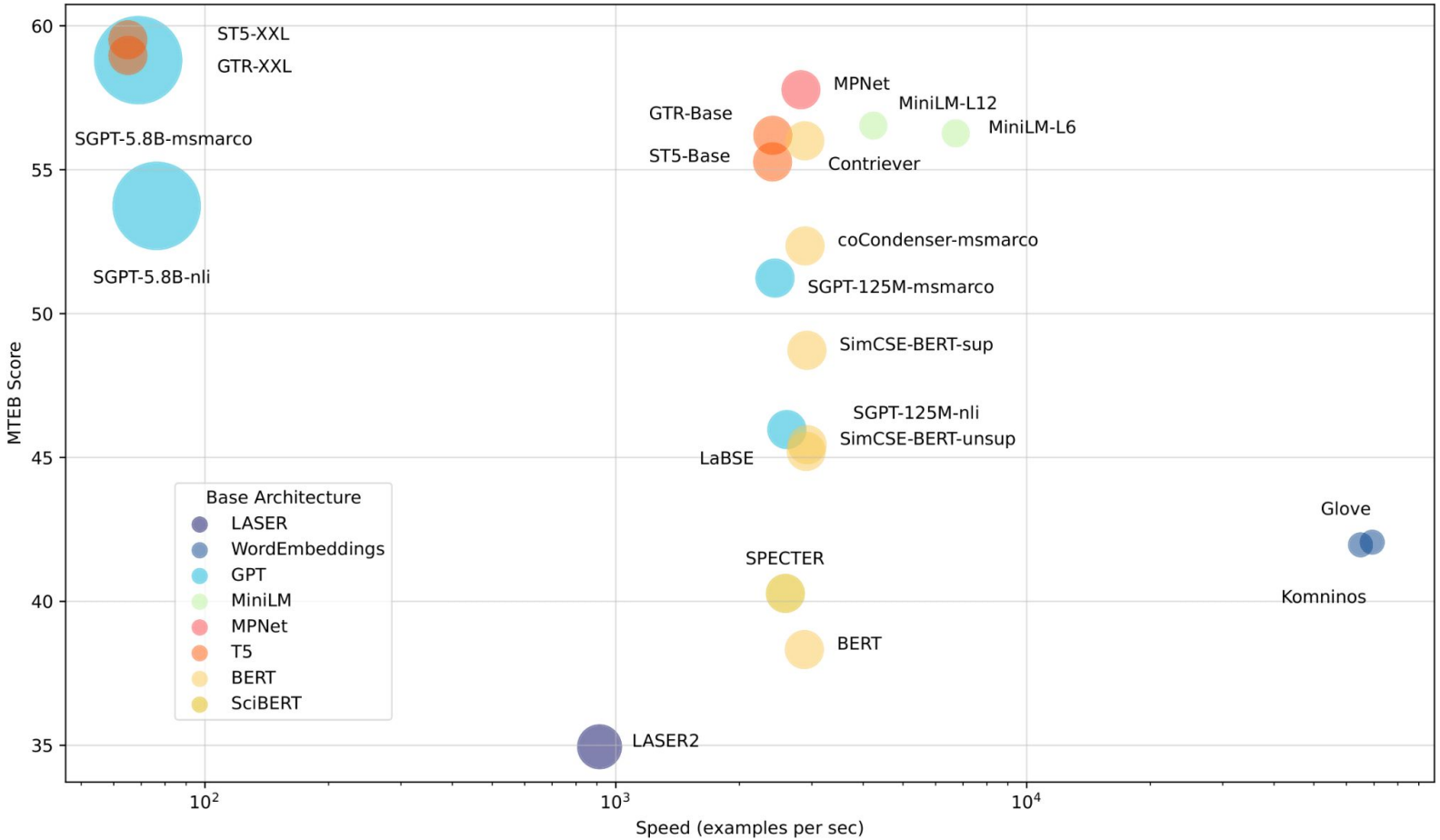
https://huggingface.co/spaces/mteb/leaderboard

Num. Datasets (→)	Class. 12	Clust. 11	PairClass. 3	Rerank. 4	Retr. 15	STS 10	Summ. 1	Avg. 56
Self-supervised methods								
Glove	57.29	27.73	70.92	43.29	21.62	61.85	28.87	41.97
Komninos	57.65	26.57	72.94	44.75	21.22	62.47	30.49	42.06
BERT	61.66	30.12	56.33	43.44	10.59	54.36	29.82	38.33
SimCSE-BERT-unsup	62.50	29.04	70.33	46.47	20.29	74.33	31.15	45.45
Supervised methods								
SimCSE-BERT-sup	67.32	33.43	73.68	47.54	21.82	79.12	31.17	48.72
coCondenser-msmarco	64.71	37.64	81.74	51.84	32.96	76.47	29.50	52.35
Contriever	66.68	41.10	82.53	53.14	41.88	76.51	30.36	56.00
SPECTER	52.37	34.06	61.37	48.10	15.88	61.02	27.66	40.28
LaBSE	62.71	29.55	78.87	48.42	18.99	70.80	31.05	45.21
LASER2	53.65	15.28	68.86	41.44	7.93	55.32	26.80	33.63
MiniLM-L6	63.06	42.35	82.37	58.04	41.95	78.90	30.81	56.26
MiniLM-L12	63.21	41.81	82.41	58.44	42.69	79.80	27.90	56.53
MiniLM-L12-multilingual	64.30	37.14	78.45	53.62	32.45	78.92	30.67	52.44
MPNet	65.07	43.69	83.04	**59.36**	43.81	80.28	27.49	57.78
MPNet-multilingual	67.91	38.40	80.81	53.80	35.34	80.73	31.57	54.71
OpenAI Ada Similarity	70.44	37.52	76.86	49.02	18.36	78.60	26.94	49.52
SGPT-125M-nli	61.46	30.95	71.78	47.56	20.90	74.71	30.26	45.97
SGPT-5.8B-nli	70.14	36.98	77.03	52.33	32.34	80.53	30.38	53.74
SGPT-125M-msmarco	60.72	35.79	75.23	50.58	37.04	73.41	29.71	51.25
SGPT-1.3B-msmarco	66.52	39.92	79.58	54.00	44.49	75.74	30.43	56.20
SGPT-2.7B-msmarco	67.13	39.83	80.65	54.67	46.54	76.83	31.03	57.17
SGPT-5.8B-msmarco	68.13	40.35	82.00	56.56	**50.25**	78.10	31.46	58.93
SGPT-BLOOM-7.1B-msmarco	66.19	38.93	81.90	55.65	48.21	77.74	**33.60**	57.59
GTR-Base	65.25	38.63	83.85	54.23	44.67	77.07	29.67	56.19
GTR-Large	67.14	41.60	85.33	55.36	47.42	78.19	29.50	58.28
GTR-XL	67.11	41.51	**86.13**	55.96	47.96	77.80	30.21	58.42
GTR-XXL	67.41	42.42	86.12	56.65	48.48	78.38	30.64	58.97
ST5-Base	69.81	40.21	85.17	53.09	33.63	81.14	31.39	55.27
ST5-Large	72.31	41.65	84.97	54.00	36.71	81.83	29.64	57.06
ST5-XL	72.84	42.34	86.06	54.71	38.47	81.66	29.91	57.87
ST5-XXL	**73.42**	**43.71**	85.06	56.43	42.24	**82.63**	30.08	**59.51**

Table 1: Average of the main metric (see Section 3.2) per task per model on MTEB English subsets.

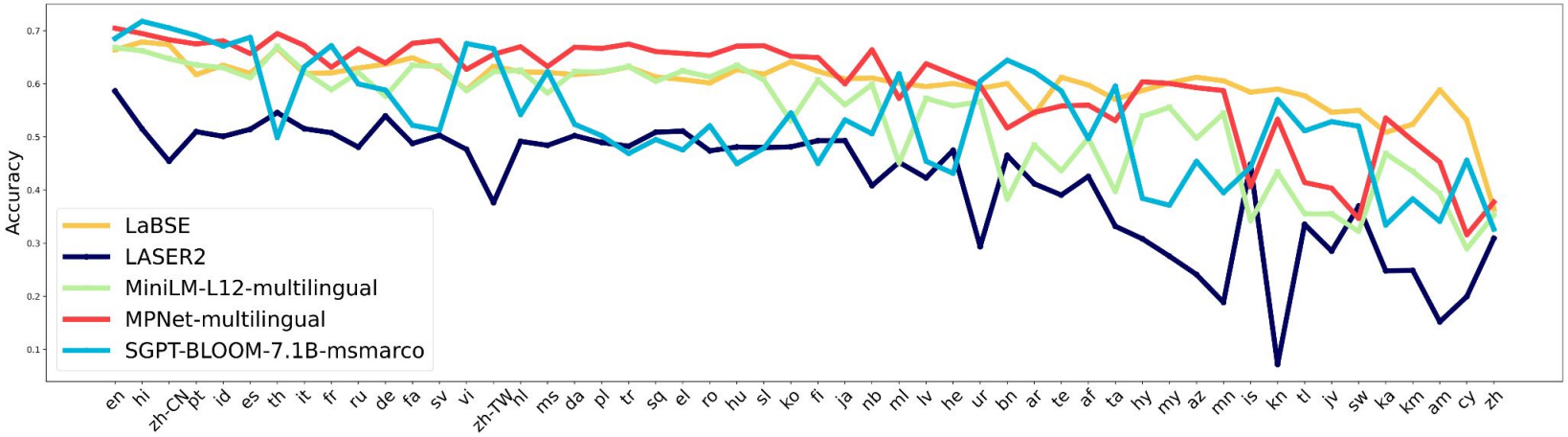


IV. EFFICIENCY

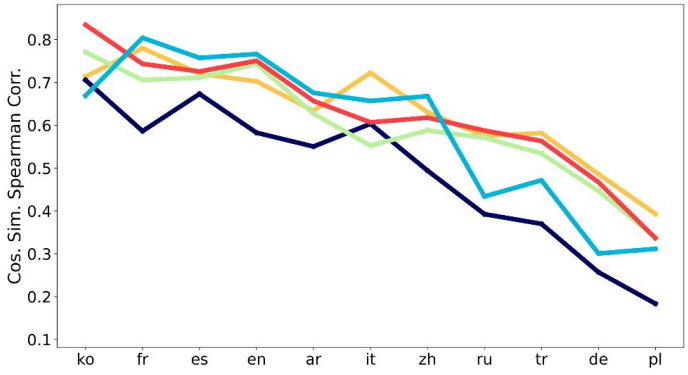


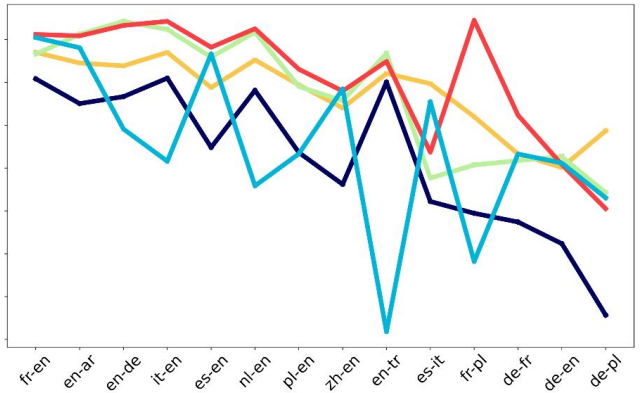
V. MULTILINGUALITY

Classification



STS





THANKS! 🤗

MTEB

Massive Text Embedding Benchmark

Niklas Muennighoff, Nouamane Tazi
Loïc Magne, Nils Reimers

Paper: arxiv.org/abs/2210.07316
Code: github.com/embeddings-benchmark/mteb
Data: hf.co/mteb
Leaderboard: hf.co/spaces/mteb/leaderboard

🤗 Hugging Face

co:here