The Russian-focused embedders' exploration: ruMTEB benchmark and Russian embedding model design

Artem Snegirev[1], Maria Tikhonova[1, 2],
Anna Maksimova[1], Alena Fenogenova[1, 2] Alexander Abramov[1]

[1]SaluteDevices, [2]HSE University **Correspondence:** artem.s.snegirev@gmail.com

Contributions

1. We publicly release **ru-en-RosBERTa** a Russian-focused text embedding model adapted for the English language.
2. We present the Russian version of MTEB and release 17 new Russian datasets for text embedding evaluation, and a public leaderboard.
3. We explore model cross-lingual transfer knowledge abilities and various model training hypotheses, which define our final training pipeline.
4. We evaluate the presented model on ruMTEB and compare its performance with a set of open-source baseline solutions.

MTEB Benchmark

The **ruMTEB benchmark** unites 23 datasets, which can be divided into 7 task categories similar to the corresponding categories in the original MTEB benchmark.

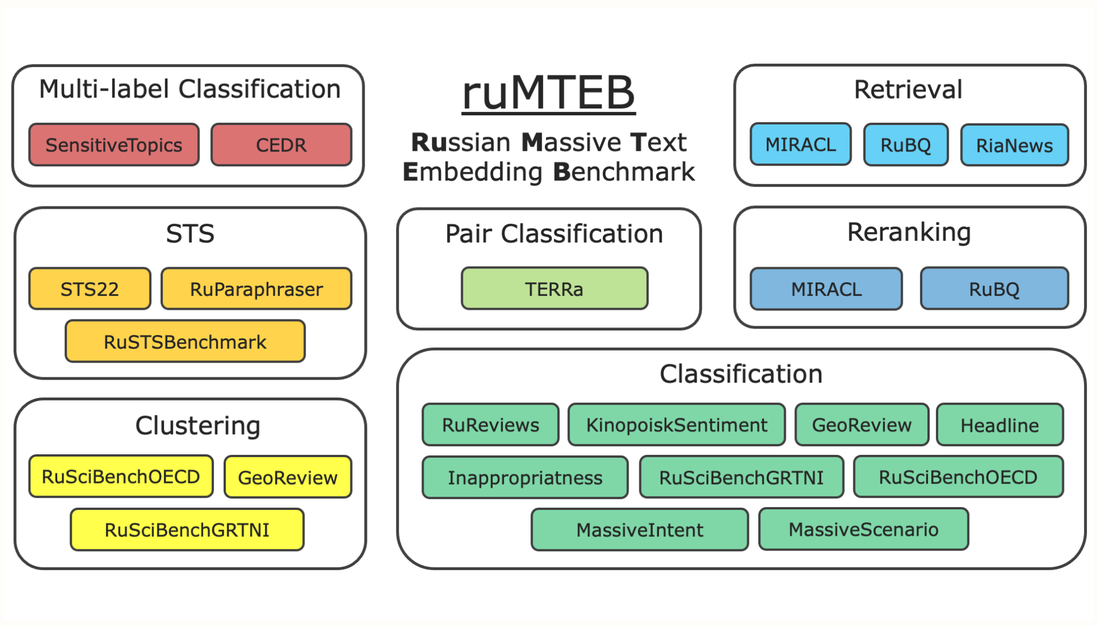

Figure 1: The scheme of the ruMTEB benchmark presenting all benchmark tasks divided into 7 task categories.

Task Category	Task name	Data origin	Train	Val	Test
Classification	GeoReviewClassification	Geo Reviews	50000	5000	5000
	HeadlineClassification	ParaPhraserPlus	36000	12000	12000
	InappropriatnessClassification	Inappropriate Sensitive Topics	4000	4000	10000
	KinopoiskSentimentClassification	Kinopoisk Movie Reviews	10500	1500	1500
	MassiveIntentClassification	MTEB	11514	2033	2974
	MassiveScenarioClassification	MTEB	11514	2033	2974
	RuReviewsClassification	RuReviews	45000	15000	15000
	RuSciBenchGRTNIClassification	RuSciBench	28476	–	2773
	RuSciBenchOECDClassification	RuSciBench	27783	–	3220
PairClassification	TERRa	TERRa	2616	307	–
MultiLabelClassification	CEDRClassification	CEDR	7529	–	1882
	SensitiveTopicsClassification	Inappropriate Sensitive topics	29178	–	2048
STS	RuSTSBenchmarkSTS	STS Benchmark	5224	1336	1264
	STS22	MTEB	–	–	265
	RuParaphraserSTS	MTEB	7227	–	1924
Clustering	GeoReviewClustering	Geo Reviews	–	–	2000
	RuSciBenchGRTNIClustering	RuSciBench	–	–	31080
	RuSciBenchOECDClustering	RuSciBench	–	–	30740
Reranking	*MIRACLReranking*	MTEB	–	44608	–
	RuBQReranking	RuBQ 2.0	–	–	1551
Retrieval	*MIRACLRetrieval*	MTEB	–	13100	–
	RiaNewsRetrieval	Ria News	–	–	10000
	RuBQRetrieval	RuBQ 2.0	–	–	2845

Table 1: The ruMTEB task outline.*Datasets from the original MTEB benchmark are in Italic; for them, the sizes of the Russian subsets are reported.

- 17 new datasets we release within the research.
- 6 datasets based on the Russian subsets from Multilingual MTEB.

Contribution to the MMTEB

Benchmark is created specifically for Russian and complex enough for modern embedding models.

References

- Muennighoff, Niklas, et al. "MTEB: Massive Text Embedding Benchmark." Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. 2023.
- Enevoldsen K. et al. MMTEB: Massive Multilingual Text Embedding Benchmark //arXiv preprint arXiv:2502.13595. – 2025.

ru-en-RoSBERTa

Training Data

- **Basic Russian Datasets**
  - SberQuAD, XNLI, OPUS/TED translations
  - Filtered web texts (news, blogs, QA forums)
- **Retrieval-Optimized**
  - Mr. TyDi & MIRACL (RU/EN)
  - Hard negatives (rank 20-100 via $mE5_{small}$)
- **Basic English Datasets**
  - MEDI corpus (30 domains)
- **Synthetic Pairs**
  - Query2doc MS-MARCO, DINO-STS
  - RuHNP/RuWANLI entailment data
  - ruT5-generated WikiOmnia samples

ru-en-RoSBERTa recipe

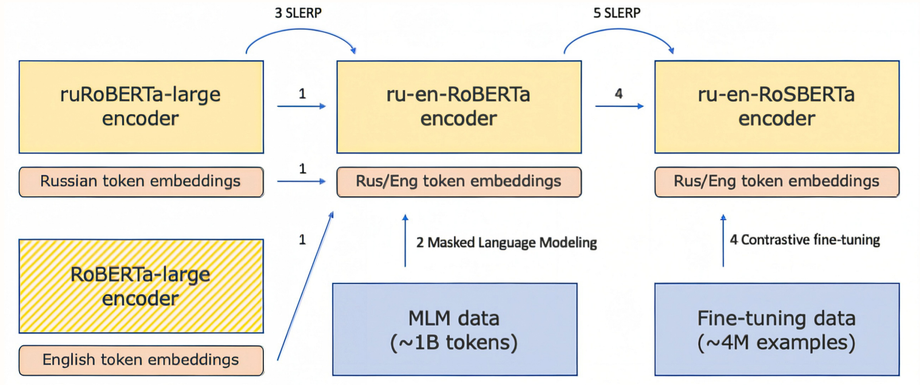

Figure 2: ru-en-RoSBERTa training recipe.

Contrastive Fine-tuning details

- InfoNCE loss (t=0.02) with CLS pooling
- Task-specific prefixes
- Stratified sampling (7 hard negatives + in-batch negatives)
- Final SLERP merge (a=0.1)

GPU resources: 8×H100 (80GB)

Takeaways and Key findings

- ru-en-RoSBERTa on par with state-of-the-art non-instruct models.
- ru-en-RoSBERTa benefits from mixture of datasets in Russian and English.
- Extended monolingual model might outperform multilingual model under limited data conditions.
- SLERP merging might be helpful for large-size encoder models.
- Prefixes are crucial for model, removal degrades performance.
- Stratified sampling is better in general especially for retrieval tasks.

Data Source	Cls.	Clust.	MultiLabelCls.	PairCls.	Rerank.	Retr.	STS	Avg.
Basic English Datasets	61.7	**56.6**	36.8	54.7	57.5	57.6	69.9	56.4
Basic Russian Datasets	60.0	54.2	38.0	56.3	60.8	61.7	72.3	57.6
Mixture	61.4	54.3	37.8	56.5	61.4	63.8	72.9	58.3
+ synthetic	**62.3**	54.6	**39.0**	59.7	62.1	64.1	73.6	59.3
+ synthetic & retrieval	62.1	53.9	**39.0**	**60.0**	**63.1**	**65.1**	**73.7**	**59.6**

Table 2: Different data sources impact.

Model name	Cls.	Clust.	MultiLabelCls.	PairCls.	Rerank.	Retr.	STS	Avg.
rubert-tiny2	52.17	39.12	29.45	51.87	30.95	8.89	61.60	42.22
$SBERT_{large\text{-}nlu\text{-}ru}$	57.24	50.44	31.87	50.17	32.81	8.51	57.21	45.35
$SBERT_{large\text{-}mt\text{-}nlu\text{-}ru}$	57.52	51.29	32.67	51.97	40.56	19.13	64.40	48.72
$mE5_{small}$	56.44	51.35	31.99	55.14	65.28	65.85	69.48	57.29
$mE5_{base}$	58.26	50.27	33.65	54.98	66.24	67.14	70.16	58.34
$mE5_{large}$	61.01	52.23	36.00	58.42	69.65	74.04	71.62	61.41
BGE-M3	60.46	52.38	34.86	60.60	69.71	**74.79**	73.68	61.58
ru-en-RoSBERTa	62.74	56.06	38.88	60.79	63.89	66.52	73.97	61.77
$mE5_{large\text{-}instruct}$	66.31	63.21	41.15	**63.89**	69.17	74.41	**74.85**	66.03
$E5_{mistral\text{-}7b\text{-}instruct}$	**69.11**	**64.24**	**42.93**	60.81	**69.96**	74.19	73.71	**67.18**

Table 3: Average model results on ruMTEB task categories.