PL-MTEB: Polish Massive Text Embedding Benchmark

Anonymous ACL submission

Abstract

In this paper, we introduce the Polish Massive Text Embedding Benchmark (PL-MTEB), a comprehensive benchmark for text embeddings in Polish. The PL-MTEB consists of 30 diverse NLP tasks from 5 task types, specifically classification, clustering, pair classification, information retrieval, and semantic textual similarity. As part of our work, We have verified the quality of the datasets available for the Polish language and prepared two new datasets, which were used to create four clustering tasks. We evaluated 30 publicly available text embedding models, including Polish and multilingual models. We made the prepared datasets, the source code for evaluation and the obtained results available to the public at [anonymized_link].

1 Introduction

Text embeddings are used in many NLP tasks, including document clustering (Aggarwal and Zhai, 2012), semantic search (Huang et al., 2020), question answering (Karpukhin et al., 2020) or classification (Muennighoff et al., 2023). In many cases, they are fundamental elements of the created systems and significantly impact their performance. Therefore, it is important to select the appropriate embedding model based on the results of its evaluation. Most often, evaluation is conducted on individual tasks using a limited set of datasets, leaving the open question of how such embedding models would work for other tasks. To solve this problem, (Muennighoff et al., 2023) created a Massive Text Embedding Benchmark (MTEB). MTEB provides a simple and clear way to examine how the model behaves for different types of tasks. Most of the tasks in MTEB were based on English-language datasets, and only a few were multilingual, making it impossible to do a good comparison of models for languages other than English. Therefore, extensions to MTEB with language-specific task sets have begun to appear, among which are C-MTEB (Xiao et al., 2024) for Chinese, MTEB for French (Ciancone et al., 2024), FaMTEB (Zinvandi et al., 2025) for Persian, MTEB-NL (Banar et al., 2025) for Dutch, ruMTEB (Snegirev et al., 2025) for Russian, VN-MTEB (Pham et al., 2025) for Vietnamese, TR-MTEB (Baysan and Gungor, 2025) for Turkish, SEB (Enevoldsen et al., 2024) for Scandinavian languages (Danish, Norwegian, Swedish), ArabicMTEB (Bhatia et al., 2025) for Arabic languages and AfriMTEB (Uemura et al., 2025) for African languages. In addition, the MMTEB (Enevoldsen et al., 2025) initiative was launched, with the aim of expanding MTEB to include new languages. In this work, we follow this path by introducing PL-MTEB (Polish Massive Text Embedding Benchmark), a comprehensive benchmark for text embeddings for Polish. Below we highlight the main contributions of this work:

- Introduction of PL-MTEB: a comprehensive benchmark consisting of 30 tasks from 5 groups (classification, clustering, pair classification, retrieval, and semantic textual similarity), designed to evaluate text embeddings for the Polish language.
- Extension of MTEB with 12 new tasks based on Polish datasets, including data quality verification.
- Preparation of two new datasets: PLSC (Polish Library of Science Corpus) and Wikinews-PL. The collections were used as a basis for proposing four new tasks for clustering.
- Evaluation of 30 models (12 for Polish and 18 multilingual) with collection of results.
- Integration with MTEB and public release of source code, all experimental results and prepared datasets.

2 Related work

2.1 Benchmarks

GLUE (Wang et al., 2018) or SuperGLUE (Wang et al., 2019) are well-known benchmarks for tracking NLP progress. They are mainly designed to compare natural language understanding systems. However, they are unsuitable for evaluating text embeddings, so dedicated benchmarks like SentEval (Conneau and Kiela, 2018) or BEIR (Thakur et al., 2021) have emerged. MTEB (Muennighoff et al., 2023) incorporates the above benchmarks, creating an accessible evaluation framework. In the case of Polish, benchmarks similar to (Super)GLUE are KLEJ (Rybak et al., 2020) and LEPISZCZE (Augustyniak et al., 2022).

To this moment, in most cases, text embeddings evaluation for Polish language was performed on individual tasks. (Krasnowska-Kieraś and Wróblewska, 2019) evaluated text embeddings on a single dataset for textual relatedness. (Dadas et al., 2020a), in their evaluation, used 3 task types (classification, textual entailment, and semantic relatedness), where only classification consisted of more than one task. (Dadas, 2022) extended this evaluation by adding 3 more tasks, one of each type. In the area of information retrieval, two benchmarks for Polish have appeared recently. The first is BEIR-PL (Wojtasik et al., 2024), which is the Polish equivalent of BEIR (Thakur et al., 2021). The second is PIRB (Dadas et al., 2024), a large benchmark consisting of 41 tasks.

2.2 Embedding Models

A few years ago, the standard method of creating text embeddings was to calculate arithmetic or weighted averages of the vectors of all words in the text. These vectors were obtained using word embedding models such as Word2Vec (Mikolov et al., 2013b,a), GloVe (Pennington et al., 2014) or FastText (Bojanowski et al., 2017). The main disadvantage of these methods was the lack of context awareness. The emergence of the Transformer (Vaswani et al., 2017) architecture, introducing context awareness through the use of the self-attention mechanism, forms the foundation of most recent embedding models. (Reimers and Gurevych, 2019) have shown that additional fine-tuning of a network composed of two transformer models leads to a model that produces high-quality sentence embeddings. Further development of the field is mainly models that use contrastive loss objective, among which we can include: SimCSE (Gao et al., 2021), TSDAE (Wang et al., 2021), GTR (Ni et al., 2022), SGPT (Muennighoff, 2022), E5 (Wang et al., 2022) or BGE (Xiao et al., 2024). With the rapid development of large language models, new text embedding methods based on them have begun to be made available. Among them, the following can be distinguished: KaLM series (Hu et al., 2025), Qwen3-Embedding (Zhang et al., 2025) and BGE-gemma2 (Xiao et al., 2024; Chen et al., 2024).

3 The PL-MTEB Benchmark

3.1 Task Types and Metrics

The benchmark consists of the following 5 task types:

Classification The classification task is to predict label based on input embedding using previously trained logistic regression classifier. Both the train and test set are embedded with the provided model. The accuracy is used as the main metric.

Clustering Given a set of sentences or paragraphs, the goal of clustering is to group them into meaningful clusters. A mini-batch k-means model with batch size 500 and k equal to the number of different labels is trained on the embedded texts. The model is scored using v-measure (Rosenberg and Hirschberg, 2007).

Pair Classification Having a pair of embedded texts, predict their relationship as binary label based on the similarity between them. The average precision score based on cosine similarity is the main metric.

Retrieval The retrieval task is presented with corpus, queries and a mapping for each query to relevant documents from the corpus. The provided model is used to embed all queries and all corpus documents. The goal is to find relevant documents based on query. Normalized Discounted Cumulative Gain at 10 (nDCG@10) serves as the main metric.

Semantic Textual Similarity (STS) Given a pair of sentences, the goal is to measure their correlation using cosine similarity score between their embeddings. Spearman correlation based on cosine similarity is used as the main metric.

3.2 Tasks

Figure 1 provides an overview of tasks available in PL-MTEB. Each task belongs to one of three groups. In the first group are tasks in Polish or multilingual tasks containing a sub-task in Polish

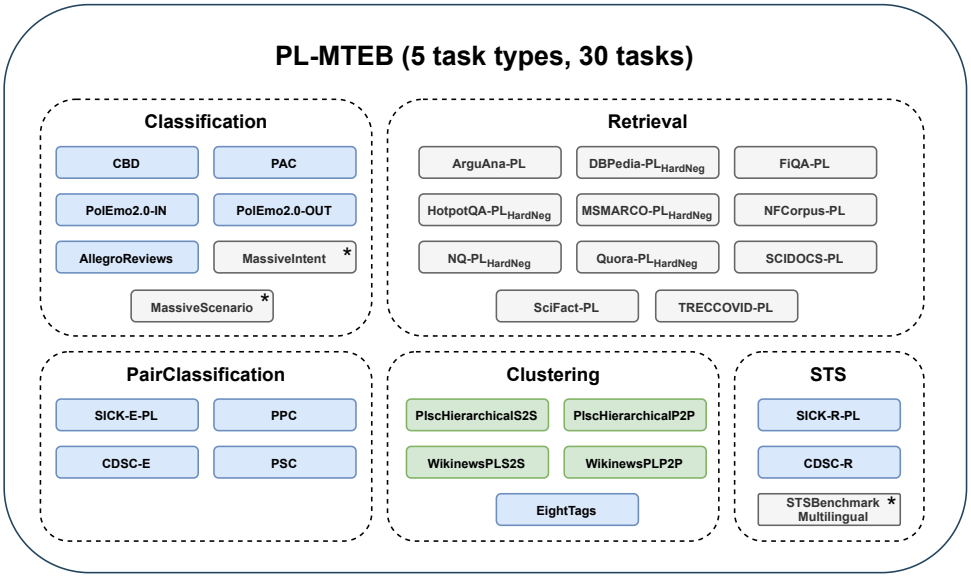

Figure 1: An overview of tasks included in PL-MTEB. The tasks with gray background are tasks in Polish that are already in MTEB (those marked with an asterisk are multilingual tasks from which we have selected Polish subtasks). The tasks marked in blue are tasks prepared in this work based on existing datasets. The green tasks were prepared on the basis of newly created datasets.

added to MTEB by other contributors. Mainly these are retrieval tasks that are part of BEIR-PL (Wojtasik et al., 2024). In the second group are tasks added by us based on existing datasets. Most of them come from the publications presented in subsection 2.1. In the third group are tasks added by us based on newly created datasets.

3.2.1 New datasets

PLSC (Polish Library of Science Corpus) is a dataset based on Library of Science[1], an open metadata repository about scientific publications. The data available there is under the CC0 license. The corpus comprises more than 100K records containing titles and abstracts of publications assigned to 8 scientific fields and 44 scientific disciplines. This collection was used to prepare two clustering tasks: PlscHierarchicalS2S and PlscHierarchicalP2P[2]. The tasks were hierarchical, i.e., first there was clustering by scientific fields, then by scientific disciplines, and the results were averaged. For performance reasons, the number of records has been limited to 2,048, in accordance with MMTEB (Enevoldsen et al., 2025) assumptions.

Wikinews-PL is a dataset of articles from the Polish version of the Wikinews portal[3] . Each article is assigned to one or more categories among the following: politics, economy, disasters, culture and entertainment, science, law and crime, sports, society and technology. Articles available there are under the CC-BY-2.5 license. The collection we downloaded consists of 15,196 articles and was used to prepare two clustering tasks: WikinewsPLS2S and WikinewsPLP2P. The number of records has also been limited to 2,048.

All of the tasks we have added are based on datasets under open licenses and are publicly available on the Hugging Face Hub[4].

3.2.2 Data Quality

During the process of creating tasks based on existing data sets, we verified their quality. First, we removed examples that were empty strings and shorter than three words. Next, we verified the labels and scores, i.e., if there were close duplicates[5] in the corpus with different labels or differing by at least 0.5 scores, they were removed. The next step was deduplication at the split level, where exact

[1] https://bibliotekanauki.pl/

[2] This is inspired by tasks from MTEB, such as ArxivS2S and ArxivP2P. S2S (sentence to sentence) and P2P (paragraph to paragraph) mean that a sentence/paragraph is compared with another sentence/paragraph, where the paragraph is in the general longer fragment of text, e.g., title + abstract.

[3] https://pl.wikinews.org

[4] https://huggingface.co/datasets

[5] To detect close duplicates, texts were normalized by converting them to lowercase and removing spaces.

Task	Reference	Test samples	Domains	Dataset Licence
Classification				
CBD	Ptaszynski et al. (2019)	999	Written, Social	BSD-3-CLAUSE
PolEmo2	Kocoń et al. (2019)	722	Written, Social	CC-BY-SA-4.0
PolEmo2	Kocoń et al. (2019)	493	Written, Social	CC-BY-SA-4.0
AllegroReviews	Rybak et al. (2020)	983	Reviews	CC-BY-SA-4.0
PAC	Augustyniak et al. (2022)	3,395	Legal, Written	CC-BY-NC-SA-4.0
MassiveIntent	FitzGerald et al. (2022)	2,974	Spoken	APACHE-2.0
MassiveScenario	FitzGerald et al. (2022)	2,974	Spoken	APACHE-2.0
Clustering				
EightTags	Dadas et al. (2020a)	2,048	Social, Written	GPL-3.0
PlscHierarchicalS2S	PL-MTEB	2,048	Academic, Written	CC0-1.0
PlscHierarchicalP2P	PL-MTEB	2,048	Academic, Written	CC0-1.0
WikinewsPlS2S	PL-MTEB	2,048	News	CC-BY-4.0
WikinewsPlP2P	PL-MTEB	2,048	News	CC-BY-4.0
Pair Classification				
SICK-E-PL	Dadas et al. (2020a)	4,874	Reviews	CC-BY-NC-SA-3.0
CDSC-E	Wróblewska and Krasnowska-Kieraś (2017)	998	Written	CC-BY-NC-SA-4.0
PSC	Ogrodniczuk and Kopeć (2014)	1,074	News, Written	CC-BY-3.0
PPC	Dadas (2022)	1,000	Fiction, Non-fiction, Web, Written, Spoken, Social, News	GPL-3.0
Retrieval				
ArguAna-PL	Wojtasik et al. (2024)	1,406 / 8,674	Medical, Written	CC-BY-SA-4.0
DBPedia-PLHardNeg	MTEB / Wojtasik et al. (2024)	400 / 88,542	Written, Encyclopaedic	MIT
FiQA-PL	Wojtasik et al. (2024)	648 / 57,638	Written, Financial	NOT SPECIFIED
HotpotQA-PLHardNeg	MTEB / Wojtasik et al. (2024)	1,000 / 212,774	Web, Written	CC-BY-SA-4.0
MSMARCO-PLHardNeg	MTEB / Wojtasik et al. (2024)	43 / 9,481	Web, Written	OWN LICENCE
NFCorpus-PL	Wojtasik et al. (2024)	323 / 3,633	Medical, Academic, Written	NOT SPECIFIED
NQ-PLHardNeg	MTEB / Wojtasik et al. (2024)	1,000 / 184,765	Written, Encyclopaedic	CC-BY-NC-SA-3.0
Quora-PLHardNeg	MTEB / Wojtasik et al. (2024)	1,000 / 172,031	Written, Web, Blog	NOT SPECIFIED
SCIDOCS-PL	Wojtasik et al. (2024)	1,000 / 25,657	Academic, Written, Non-fiction	CC-BY-SA-4.0
SciFact-PL	Wojtasik et al. (2024)	300 / 5,183	Academic, Medical, Written	NOT SPECIFIED
TRECCOVID-PL	Wojtasik et al. (2024)	50 / 171,332	Academic, Medical, Non-fiction, Written	NOT SPECIFIED
STS				
SICK-R-PL	Dadas et al. (2020a)	4,871	Web, Written	CC-BY-NC-SA-3.0
CDSC-R	Wróblewska and Krasnowska-Kieraś (2017)	998	Web, Written	CC-BY-NC-SA-4.0
STSBenchmarkMultilingual	MTEB	1,379	News, Social, Web, Spoken, Written	NOT SPECIFIED

Table 1: Tasks in PL-MTEB. The two numbers in the test samples column for an retrieval tasks represent the number of questions and the size of the corpus, respectively.

and close duplicates were removed in turn. The final step was to verify that there was no test-train leakage. As a result of this process, we obtained datasets that were then used to prepare tasks for PL-MTEB.

4 Evaluation

4.1 Models

We conducted the evaluations on dense embedding models that were trained in a supervised manner and were recently SOTA solutions. Below is a brief description of the evaluated models.

LaBSE (Feng et al., 2022) A language-agnostic BERT sentence embedding model supporting 109 languages optimized for bi-text mining tasks.

Multilingual SBERT (Reimers and Gurevych, 2019) Sentence-BERT (SBERT) is a modification of the pretrained BERT (Devlin et al., 2019) network that use siamese and triplet network structures to generate text embeddings. In our experiments we use three SBERT multilingual models: *distiluse-base-multilingual-cased-v2*, *paraphrase-multilingual-MiniLM-L12-v2*, *paraphrase-multilingual-mpnet-base-v2*, and *static-similarity-mrl-multilingual-v1*. These models were created using knowledge distillation (Reimers and Gurevych, 2020).

Multilingual E5 (Wang et al., 2022) Text encoder supporting over 100 languages, developed using two-stage training procedure. The first stage involved weakly-supervised training on a dataset of text pairs extracted from large internet corpora, such as Common Crawl. In the second stage, the model was fine-tuned in a supervised manner on several annotated datasets. We use three versions of this model: small, base, and large.

KaLM (Hu et al., 2025) Series of embedding models adapted from auto-regressive LLMs with superior training data. *HIT-TMG/KaLM-embedding-multilingual-mini-instruct-v1* was trained from Qwen/Qwen2-0.5B with massive weakly-supervised pre-training and supervised fine-tuning data.

Snowflake's Arctic (Yu et al., 2024) Multilingual embedding models trained in a multi-stage pipeline to optimize their retrieval performance.

DRAMA (Ma et al., 2025) Dense retrieval models built upon a pruned large language model back-

Model name / (# tasks)	Model size	Zero shot	Class. (7)	Clust. (5)	PairClass. (4)	Retr. (11)	STS (3)	Avg. (30)	Avg. (by type)
Multilingual									
LaBSE	471.0M	100	57.35	42.40	79.27	27.36	74.67	48.52	56.21
distiluse-base-multilingual-cased-v2	134.7M	93	48.95	38.86	79.37	24.68	75.75	45.10	53.52
paraphrase-multilingual-MiniLM-L12-v2	118.0M	93	51.39	40.68	83.40	30.40	78.68	48.91	56.91
paraphrase-multilingual-mpnet-base-v2	278.0M	93	53.23	41.34	86.21	33.33	81.13	51.14	59.05
static-similarity-mrl-multilingual-v1	108.4M	96	48.17	30.04	70.41	24.84	72.01	41.95	49.09
multilingual-e5-small	118.0M	90	52.64	43.99	81.70	46.00	78.41	55.21	60.55
multilingual-e5-base	278.0M	90	55.36	44.10	82.08	47.63	79.13	56.59	61.66
multilingual-e5-large	560.0M	90	58.53	40.60	84.57	52.43	81.41	59.06	63.51
KaLM-embedding-multilingual-mini-instruct-v1	494.0M	63	64.89	53.63	80.68	44.59	76.24	58.81	64.01
snowflake-arctic-embed-l-v2.0	568.0M	93	57.12	43.56	80.20	54.29	77.95	58.98	62.62
snowflake-arctic-embed-m-v2.0	305.0M	90	54.01	43.80	78.37	52.21	75.60	57.06	60.80
drama-base	211.8M	90	42.06	40.48	72.05	28.29	65.01	43.04	49.58
drama-large	399.8M	90	45.15	41.61	74.41	33.22	67.05	46.28	52.29
drama-1b	1.2B	90	58.46	45.11	80.60	51.49	78.21	58.61	62.77
Qwen3-Embedding-0.6B	595.8M	90	69.66	56.65	81.31	48.59	78.45	62.20	66.93
Qwen3-Embedding-4B	4.0B	90	79.3	**59.9**	86.68	56.65	85.55	69.37	73.62
Qwen3-Embedding-8B	7.6B	90	**79.87**	58.64	87.61	59.21	86.72	**70.47**	**74.41**
bge-multilingual-gemma2	9.2B	83	77.77	58.15	**89.75**	58.93	83.97	69.81	73.71
Polish									
silver-retriever-base-v1.1	124.4M	100	57.03	44.92	74.82	42.92	74.61	53.97	58.86
st-polish-paraphrase-from-mpnet	124.4M	100	57.57	44.53	87.06	38.33	82.83	54.80	62.06
st-polish-paraphrase-from-distilroberta	124.4M	100	57.71	42.71	86.96	36.16	82.63	53.70	61.23
mmlw-e5-small	117.7M	90	60.12	48.91	86.67	46.43	82.05	58.97	64.84
mmlw-e5-base	278.0M	90	47.37	37.29	59.50	53.70	49.02	49.80	49.38
mmlw-e5-large	559.9M	90	53.59	38.93	59.80	56.53	39.95	51.69	49.76
mmlw-roberta-base	124.4M	96	62.53	48.00	88.16	53.60	85.20	62.52	67.50
mmlw-roberta-large	435.0M	96	66.15	44.58	89.15	49.91	85.23	61.58	67.00
mmlw-retrieval-roberta-large	435.0M	93	63.90	45.18	88.48	57.23	84.71	63.69	67.90
mmlw-retrieval-roberta-large-v2	435.0M	80	64.62	39.08	86.53	58.35	85.64	63.09	66.84
stella-pl	1.5B	80	66.94	38.08	89.20	60.82	**86.87**	64.85	68.38
stella-pl-retrieval-8k	1.5B	80	68.14	35.42	89.56	**61.59**	86.56	64.98	68.25

Table 2: Average of the main metric per task type and overall score on PL-MTEB. The best score for a given column is marked in **bold**, and the second best is underlined.

bone and fine-tuned for efficient and generalizable multilingual text retrieval.

Qwen3-Embedding (Zhang et al., 2025) Qwen3 Embedding model series specifically designed for text embedding and ranking tasks. We evaluated three versions of this model: 0.6B, 4B, and 8B.

BGE-gemma2 (Xiao et al., 2024; Chen et al., 2024) LLM-based multilingual embedding model. It is trained on a diverse range of languages and tasks based on google/gemma-2-9b.

Silver Retriever (Rybak and Ogrodniczuk, 2024) Polish dense retrieval model trained on MAUPQA (Rybak, 2023) datasets with hard negatives mined employing a combination of heuristic rules and cross-encoders. Model was based on HerBERT language model (Mroczkowski et al., 2021).

Polish SBERT (Dadas, 2022) SBERT model trained using multilingual knowledge distillation technique and Polish-English bilingual corpus. In our experiments we use two such models: *st-polish-paraphrase-from-mpnet* and *st-polish-paraphrase-from-distilroberta*.

MMLW (Dadas et al., 2024) Another set of models trained using the knowledge distillation technique. Authors selected two groups of models as student models: pre-trained Polish RoBERTa language models (Dadas et al., 2020b) and multilingual E5 (Wang et al., 2022). As teachers, they chose English BGE (Xiao et al., 2024) models. Specifically, these models were: *mmlw-roberta-base*, *mmlw-roberta-large*, *mmlw-e5-small*, *mmlw-e5-base* and *mmlw-e5-large*. In addition, we tested two mmlw models prepared for retrieval: *mmlw-retrieval-roberta-large* and *mmlw-retrieval-roberta-large-v2*

Stella-PL (Dadas et al., 2024) Bilingual Polish-English text encoder based on stella_en_1.5B_v5, prepared similarly to mmlw model by using the knowledge distillation technique. Model *stella-pl-retrieval-8k* has expanded context.

4.2 Main Results

The results of our experiments are presented in Table 2. The best average results were achieved by the largest models based on LLMs. The best result for the entire set was achieved by **Qwen3-Embedding-8B** (70.47 / 74.42), followed by **bge-multilingual-gemma2** (69.82 / 73.71). At the same time, it should be noted that these are average results, and a more detailed look will allow other findings to be identified. No model was the best in all tasks. Models from Qwen3-Embeddings family were the best for classification and clustering tasks. For pair classification the best was **bge-multilingual-gemma2**.

In retrieval and STS Stella-PL models were the best. Additional information that should be taken into account when interpreting the results is the zero-shot column. It indicates what percentage of tasks in the benchmark are new to the model and were not included in the training. It should be noted that we are not talking about using exactly the same examples, but only training sets that are part of the same data sets. Taking this into account, all of the best-performing models had some of these types of tasks in their training sets.

5 Conclusion

In this work, we have introduced PL-MTEB, a text embedding benchmark for the Polish language consisting of 30 tasks belonging to 5 categories. We evaluated 30 models, including Polish and multilingual ones. The **Qwen3-Embedding-8B** achieved the best average results for all task types. We believe that our work will help standardize the evaluation of text embedding models for Polish. At the same time, tasks from PL-MTEB can be used by the wider international community for more accurate evaluations of multilingual embeddings. PL-MTEB is a benchmark that will be successively updated with results for new models. Given the public nature of our benchmark and the findings related to zero-shot settings, we plan to expand the benchmark to include closed tasks in the future.

The source code for evaluating new models or reproducing our experiments is available at [anonymized_link]. Datasets and public leaderboard can be found at [anonymized_link]. As PL-MTEB is part of the MTEB project, the source code of the tasks themselves and details related to the evaluation are at `https://github.com/embeddings-benchmark/mteb`.

Limitations

Long document datasets PL-MTEB includes texts of varying lengths, but most of them are not long. There are no tasks with very long texts.

Closed-source models We evaluated only publicly available models, excluding closed ones accessible via API, such as *text-embedding-3-small* from OpenAI. This was due to the limited budget of the project. We plan to include such solutions in the future.

References

Charu C. Aggarwal and ChengXiang Zhai. 2012. *A Survey of Text Clustering Algorithms*, pages 77–128. Springer US, Boston, MA.

Lukasz Augustyniak, Kamil Tagowski, Albert Sawczyn, Denis Janiak, Roman Bartusiak, Adrian Szymczak, Arkadiusz Janz, Piotr Szymański, Marcin Wątroba, Mikoł aj Morzy, Tomasz Kajdanowicz, and Maciej Piasecki. 2022. This is the way: designing and compiling LEPISZCZE, a comprehensive NLP benchmark for Polish. In *Advances in Neural Information Processing Systems*, volume 35, pages 21805–21818. Curran Associates, Inc.

Nikolay Banar, Ehsan Lotfi, Jens Van Nooten, Cristina Arhiliuc, Marija Kliocaite, and Walter Daelemans. 2025. Mteb-nl and e5-nl: Embedding benchmark and models for dutch. *Preprint*, arXiv:2509.12340.

Mehmet Selman Baysan and Tunga Gungor. 2025. TR-MTEB: A comprehensive benchmark and embedding model suite for Turkish sentence representations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8867–8887, Suzhou, China. Association for Computational Linguistics.

Gagan Bhatia, El Moatez Billah Nagoudi, Abdellah El Mekki, Fakhraddin Alwajih, and Muhammad Abdul-Mageed. 2025. Swan and ArabicMTEB: Dialect-aware, Arabic-centric, cross-lingual, and cross-cultural embedding models and benchmarks. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4654–4670, Albuquerque, New Mexico. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Mathieu Ciancone, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. 2024. Extending the massive text embedding benchmark to french. *Preprint*, arXiv:2405.20468.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Slawomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2020a. Evaluation of Sentence Representations in Polish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1674–1680, Marseille, France. European Language Resources Association.

Sławomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2020b. Pre-training polish transformer-based language models at scale. In *Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part II 19*, pages 301–314. Springer.

Slawomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2024. PIRB: A comprehensive benchmark of Polish dense and hybrid text retrieval methods. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12761–12774, Torino, Italia. ELRA and ICCL.

Sławomir Dadas. 2022. Training Effective Neural Sentence Encoders from Automatically Mined Paraphrases. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 371–378.

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.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzeminski, {Genta Indra} Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, and 37 others. 2025. Mmteb: Massive multilingual text embedding benchmark. In *13th International Conference on Learning Representations, ICLR 2025*, pages 102004–102060. International Conference on Learning Representations.

Kenneth Enevoldsen, Márton Kardos, Niklas Muennighoff, and Kristoffer Nielbo. 2024. The scandinavian embedding benchmarks: Comprehensive assessment of multilingual and monolingual text embedding. In *Advances in Neural Information Processing Systems*.

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.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *Preprint*, arXiv:2204.08582.

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.

Xinshuo Hu, Zifei Shan, Xinping Zhao, Zetian Sun, Zhenyu Liu, Dongfang Li, Shaolin Ye, Xinyuan Wei, Qian Chen, Baotian Hu, and 1 others. 2025. Kalm-embedding: Superior training data brings a stronger embedding model. *arXiv preprint arXiv:2501.01028*.

Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 2553–2561, New York, NY, USA. Association for Computing Machinery.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Jan Kocoń, Piotr Miłkowski, and Monika Zaśko-Zielińska. 2019. Multi-level sentiment analysis of PolEmo 2.0: Extended corpus of multi-domain consumer reviews. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 980–991, Hong Kong, China. Association for Computational Linguistics.

Katarzyna Krasnowska-Kieraś and Alina Wróblewska. 2019. Empirical Linguistic Study of Sentence Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5729–5739, Florence, Italy. Association for Computational Linguistics.

Xueguang Ma, Xi Victoria Lin, Barlas Oguz, Jimmy Lin, Wen-tau Yih, and Xilun Chen. 2025. DRAMA: Diverse augmentation from large language models to smaller dense retrievers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30170–30186, Vienna, Austria. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tomas Mikolov, Kai Chen, and Greg Corrado ands Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. Herbert: Efficiently pretrained transformer-based language model for polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *Preprint*, arXiv:2202.08904.

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.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maciej Ogrodniczuk and Mateusz Kopeć. 2014. The Polish Summaries Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Loc Pham, Tung Luu, Thu Vo, Minh Nguyen, and Viet Hoang. 2025. Vn-mteb: Vietnamese massive text embedding benchmark. *Preprint*, arXiv:2507.21500.

Michal Ptaszynski, Agata Pieciukiewicz, and Paweł Dybała. 2019. Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter. *Proceedings of the PolEval 2019 Workshop*, page 89.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Andrew Rosenberg and Julia Hirschberg. 2007. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic. Association for Computational Linguistics.

Piotr Rybak. 2023. Maupqa: Massive automatically-created polish question answering dataset. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 11–16.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. KLEJ: Comprehensive Benchmark for Polish Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1191–1201, Online. Association for Computational Linguistics.

Piotr Rybak and Maciej Ogrodniczuk. 2024. Silver retriever: Advancing neural passage retrieval for Polish question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14826–14831, Torino, Italia. ELRA and ICCL.

Artem Snegirev, Maria Tikhonova, Maksimova Anna, Alena Fenogenova, and Aleksandr Abramov. 2025. The Russian-focused embedders' exploration: ruMTEB benchmark and Russian embedding model design. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 236–254, Albuquerque, New Mexico. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Kosei Uemura, Miaoran Zhang, and David Ifeoluwa Adelani. 2025. Afrimteb and afrie5: Benchmarking and adapting text embedding models for african languages. *Preprint*, arXiv:2510.23896.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*. Curran Associates Inc., Red Hook, NY, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. 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.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text Embeddings by Weakly-Supervised Contrastive Pre-training. *Preprint*, arXiv:2212.03533.

Konrad Wojtasik, Kacper Wołowiec, Vadim Shishkin, Arkadiusz Janz, and Maciej Piasecki. 2024. BEIR-PL: Zero shot information retrieval benchmark for the Polish language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2149–2160, Torino, Italia. ELRA and ICCL.

Alina Wróblewska and Katarzyna Krasnowska-Kieraś. 2017. Polish evaluation dataset for compositional distributional semantics models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 784–792.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. 2024. Arctic-embed 2.0: Multilingual retrieval without compromise. *Preprint*, arXiv:2412.04506.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Erfan Zinvandi, Morteza Alikhani, Mehran Sarmadi, Zahra Pourbahman, Sepehr Arvin, Reza Kazemi, and Arash Amini. 2025. FaMTEB: Massive text embedding benchmark in Persian language. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11441–11468, Suzhou, China. Association for Computational Linguistics.

A Tasks Descriptions

A.1 Classification

CBD (Ptaszynski et al., 2019) The Cyberbullying Detection task, where the goal is to predict if tweet contains a cyberbullying content.

PAC (Augustyniak et al., 2022) Polish Abusive Clauses Dataset used to formulate binary classification task of detecting abusive clauses.

PolEmo2.0-IN and **PolEmo2.0-OUT** (Kocoń et al., 2019) Based on a collection of Polish online reviews from four domains: medicine, hotels, products and school. The PolEmo2.0-IN task is to predict the sentiment of in-domain (medicine and hotels) reviews. The PolEmo2.0-OUT task is to predict the sentiment of out-of-domain (products and school) reviews using models train on reviews from medicine and hotels domains.

MassiveIntent and **MassiveScenario** (FitzGerald et al., 2022) The tasks include intent and scenario detection from the content of utterances addressed to Amazon's Alexa virtual assistant. They are based on a multilingual dataset with 51 available languages, of which we use only Polish-language subset. The tasks were already in MTEB.

AllegroReviews (Rybak et al., 2020) Based on a Polish dataset for sentiment classification on reviews from e-commerce marketplace Allegro. The task is to predict a rating ranging from 1 to 5.

A.2 Clustering

EightTags (original name 8Tags) (Dadas et al., 2020a) Clustering of headlines from social media posts in Polish belonging to 8 categories: film, history, food, medicine, motorization, work, sport and technology.

PlscHierarchicalS2S and **PlscHierarchicalP2P** Tasks involve clustering publication titles and titles with abstracts, respectively, first in terms of their scientific field and than by scientific disciplines.
WikinewsPLS2S and **WikinewsPLP2P** Tasks involve clustering Wikinews article titles and titles with texts, respectively, in terms of category.

A.3 Pair Classification

SICK-E-PL (Dadas et al., 2020a) The binary variant of textual entailment task based on the Polish version of Sentences Involving Compositional Knowledge (SICK) (Marelli et al., 2014) dataset, where labels 'neutral' and 'contradiction' was merged to create one 'not entailed' class.
CDSC-E (Wróblewska and Krasnowska-Kieraś, 2017) The binary variant of textual entailment task based on Compositional Distributional Semantics Corpus, where labels 'neutral' and 'contradiction' was merged to create one 'not entailed' class.
PPC (Dadas, 2022) A task to detect whether a given sentence is a paraphrase of another. Based on a Polish Paraphrase Corpus, class 'exact paraphrase' and 'close paraphrase' are merged.
PSC (Ogrodniczuk and Kopeć, 2014) The task is to detect whether two summaries relate to the same article. Base on The Polish Summaries Corpus.

A.4 Retrieval

The vast majority of retrieval tasks are from BEIR-PL (Wojtasik et al., 2024), which was created by automatic translating dataset from BEIR (Thakur et al., 2021) to Polish language.
ArguAna-PL Retrieving the best counterargument to a given argument.
DBPedia-PL Searching for entities in the DBpedia knowledge base.
FiQA-PL Retrieving relevant documents from financial domain to a given query.
HotpotQA-PL A question answering task which requires reasoning over multiple paragraphs (multi-hop) and Wikipedia articles are the information source.
MSMARCO-PL A question answering task based on Bing questions and human generated answers.
NFCorpus-PL Retrieving relevant documents from NutrionFacts (medicine domain) to a given query.
NQ-PL A question answering task where the questions are from a Google search engine and the answers are annotated by a human based on Wikipedia articles.
Quora-PL Task is based on questions that are marked as duplicates on the Quora platform. Given a question, find other (duplicate) questions.
SCIDOCS-PL Citation prediction task, where the goal is to get cited scientific articles based on the title of the article that cites them.
SciFact-PL Verifing scientific claims using evidence from the research literature containing scientific paper abstracts.
TRECCOVID-PL Retrieving relevant scientific articles related to COVID-19 based on a given query.

A.5 Semantic Textual Similarity (STS)

SICK-R-PL (Dadas et al., 2020a) Textual relatedness task based on Polish version of Sentences Involving Compositional Knowledge (SICK) (Marelli et al., 2014) dataset.
CDSC-R (Wróblewska and Krasnowska-Kieraś, 2017) Textual relatedness task based on Compositional Distributional Semantics Corpus.
STSBenchmarkMultilingual Semantic Textual Similarity Benchmark (STSbenchmark) dataset, translated using DeepL API. Source of the dataset: `https://github.com/PhilipMay/stsb-multi-mt`.
We use only Polish-language subset. The task was already in MTEB.

B Models

Table 3 contains references to the evaluated models.

C Results

Detailed results for each type of task are presented in Tables 4–8. These results show, among other things, that most models were trained on retrieval data, which is why the zero-shot score for these models is less than 100%.

Name in Paper	HF Name
LaBSE	sentence-transformers/LaBSE
distiluse-base-multilingual-cased-v2	sentence-transformers/distiluse-base-multilingual-cased-v2
paraphrase-multilingual-MiniLM-L12-v2	sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
paraphrase-multilingual-mpnet-base-v2	sentence-transformers/paraphrase-multilingual-mpnet-base-v2
static-similarity-mrl-multilingual-v1	sentence-transformers/static-similarity-mrl-multilingual-v1
multilingual-e5-small	intfloat/multilingual-e5-small
multilingual-e5-base	intfloat/multilingual-e5-base
multilingual-e5-large	intfloat/multilingual-e5-large
KaLM-embedding-multilingual-mini-instruct-v1	HIT-TMG/KaLM-embedding-multilingual-mini-instruct-v1
snowflake-arctic-embed-l-v2.0	Snowflake/snowflake-arctic-embed-l-v2.0
snowflake-arctic-embed-m-v2.0	Snowflake/snowflake-arctic-embed-m-v2.0
drama-base	facebook/drama-base
drama-large	facebook/drama-large
drama-1b	facebook/drama-1b
Qwen3-Embedding-0.6B	Qwen/Qwen3-Embedding-0.6B
Qwen3-Embedding-4B	Qwen/Qwen3-Embedding-4B
Qwen3-Embedding-8B	Qwen/Qwen3-Embedding-8B
bge-multilingual-gemma2	BAAI/bge-multilingual-gemma2
silver-retriever-base-v1.1	ipipan/silver-retriever-base-v1.1
st-polish-paraphrase-from-mpnet	sdadas/st-polish-paraphrase-from-mpnet
st-polish-paraphrase-from-distilroberta	sdadas/st-polish-paraphrase-from-distilroberta
mmlw-e5-small	sdadas/mmlw-e5-small
mmlw-e5-base	sdadas/mmlw-e5-base
mmlw-e5-large	sdadas/mmlw-e5-large
mmlw-roberta-base	sdadas/mmlw-roberta-base
mmlw-roberta-large	sdadas/mmlw-roberta-large
mmlw-retrieval-roberta-large	sdadas/mmlw-retrieval-roberta-large
mmlw-retrieval-roberta-large-v2	sdadas/mmlw-retrieval-roberta-large-v2
stella-pl	sdadas/stella-pl
stella-pl-retrieval-8k	sdadas/stella-pl-retrieval-8k

Table 3: Model names as referenced in the paper, and corresponding Hugging Face Hub identifiers.

Model name	Zero shot	CBD	PolEmo2.0-IN	PolEmo2.0-OUT	AllegroReviews	PAC	MassiveIntent	MassiveScenario	Avg.
LaBSE	100	64.69	64.56	47.24	35.44	65.58	59.83	64.12	57.35
distiluse-base-multilingual-cased-v2	100	51.94	51.09	32.29	28.69	64.63	52.85	61.15	48.95
paraphrase-multilingual-MiniLM-L12-v2	100	53.56	59.24	28.34	31.10	62.77	59.54	65.16	51.39
paraphrase-multilingual-mpnet-base-v2	100	57.77	62.78	19.76	36.19	62.48	64.75	68.87	53.23
static-similarity-mrl-multilingual-v1	100	54.19	53.38	38.40	26.40	56.63	53.78	54.40	48.17
multilingual-e5-small	100	58.22	58.05	24.28	35.35	**71.03**	57.96	63.58	52.64
multilingual-e5-base	100	57.35	58.88	35.80	37.76	70.09	61.82	65.79	55.36
multilingual-e5-large	100	61.50	65.58	38.17	39.21	70.48	66.07	68.67	58.53
KaLM-embedding-multilingual-mini-instruct-v1	71	61.35	78.61	61.36	56.30	62.13	62.49	71.99	64.89
snowflake-arctic-embed-l-v2.0	100	65.22	62.51	34.71	31.87	64.96	68.22	72.38	57.12
snowflake-arctic-embed-m-v2.0	100	62.52	58.20	28.17	29.89	64.97	64.84	69.51	54.01
drama-base	100	49.38	52.59	23.59	28.12	58.41	37.31	44.99	42.06
drama-large	100	53.61	52.99	24.14	28.73	60.23	44.22	52.12	45.15
drama-1b	100	59.45	68.31	47.38	40.62	63.43	62.72	67.29	58.46
Qwen3-Embedding-0.6B	100	63.42	87.42	71.74	59.88	61.60	70.38	73.18	69.66
Qwen3-Embedding-4B	100	81.41	90.37	77.73	68.95	69.89	81.24	85.5	79.3
Qwen3-Embedding-8B	100	**83.71**	91.29	**79.41**	**69.37**	65.22	**83.11**	**86.96**	**79.87**
bge-multilingual-gemma2	100	82.6	**91.63**	78.4	64.53	66.16	79.52	81.57	77.77
silver-retriever-base-v1.1	100	63.36	62.60	43.31	33.57	61.68	66.45	68.27	57.03
st-polish-paraphrase-from-mpnet	100	67.30	67.83	31.62	35.35	63.13	66.04	71.75	57.57
st-polish-paraphrase-from-distilroberta	100	64.96	66.02	40.97	33.27	63.46	65.09	70.17	57.71
mmlw-e5-small	100	60.87	70.11	47.24	35.23	64.82	69.66	72.90	60.12
mmlw-e5-base	100	52.93	47.40	34.50	25.13	62.82	53.09	55.74	47.37
mmlw-e5-large	100	50.72	63.60	42.74	34.65	65.83	56.23	61.36	53.59
mmlw-roberta-base	100	63.15	73.03	47.81	39.82	65.86	72.55	75.50	62.53
mmlw-roberta-large	100	64.44	77.58	55.60	47.24	65.33	75.13	77.74	66.15
mmlw-retrieval-roberta-large	100	65.13	70.50	52.68	41.00	63.67	76.14	78.17	63.90
mmlw-retrieval-roberta-large-v2	100	62.89	75.98	55.15	40.92	67.84	72.50	77.07	64.62
stella-pl	100	65.19	82.05	60.28	48.07	62.35	73.50	77.16	66.94
stella-pl-retrieval-8k	100	67.17	82.80	64.00	48.48	63.51	74.02	77.00	68.14

Table 4: Evaluation results on classification tasks using accuracy metric. The best score for a given column is marked in **bold**, and the second best is underlined.

Model name	Zero shot	EightTags	PlscHierarchicalS2S	PlscHierarchicalP2P	WikinewsPLS2S	WikinewsPLP2P	Avg.
LaBSE	100	26.11	48.45	57.06	35.40	44.99	42.40
distiluse-base-multilingual-cased-v2	100	26.90	41.98	51.42	31.78	42.20	38.86
paraphrase-multilingual-MiniLM-L12-v2	100	26.14	47.64	54.75	30.54	44.33	40.68
paraphrase-multilingual-mpnet-base-v2	100	29.41	48.89	51.52	32.81	44.08	41.34
static-similarity-mrl-multilingual-v1	100	16.93	37.57	46.63	19.01	30.08	30.04
multilingual-e5-small	100	30.21	49.83	55.88	41.48	42.55	43.99
multilingual-e5-base	100	31.17	49.67	53.63	40.94	45.11	44.10
multilingual-e5-large	100	27.18	50.49	53.74	31.13	40.46	40.60
KaLM-embedding-multilingual-mini-instruct-v1	100	38.84	52.63	60.89	55.67	60.14	53.63
snowflake-arctic-embed-l-v2.0	100	33.47	51.64	55.52	38.00	39.17	43.56
snowflake-arctic-embed-m-v2.0	100	30.12	49.94	54.42	41.75	42.77	43.80
drama-base	100	24.90	47.28	53.22	28.22	48.79	40.48
drama-large	100	26.98	48.98	53.48	29.40	49.21	41.61
drama-1b	100	33.18	51.56	54.76	37.49	48.54	45.11
Qwen3-Embedding-0.6B	100	46.65	55.47	62.56	59.21	59.36	56.65
Qwen3-Embedding-4B	100	**62.3**	56.57	60.69	**59.62**	**60.3**	**59.9**
Qwen3-Embedding-8B	100	60.4	56.19	61.22	55.74	59.63	58.64
bge-multilingual-gemma2	100	59.27	**58.68**	**62.95**	54.01	55.82	58.15
silver-retriever-base-v1.1	100	32.18	49.19	56.88	39.68	46.67	44.92
st-polish-paraphrase-from-mpnet	100	31.30	49.31	56.94	37.65	47.43	44.53
st-polish-paraphrase-from-distilroberta	100	30.40	47.47	55.78	35.30	44.58	42.71
mmlw-e5-small	100	32.28	52.40	56.96	44.66	58.25	48.91
mmlw-e5-base	100	23.72	44.23	53.88	26.47	38.14	37.29
mmlw-e5-large	100	27.93	45.04	55.39	27.30	39.01	38.93
mmlw-roberta-base	100	31.61	51.02	58.35	46.11	52.89	48.00
mmlw-roberta-large	100	33.35	53.66	56.97	34.93	43.98	44.58
mmlw-retrieval-roberta-large	100	31.79	51.66	55.47	41.22	45.74	45.18
mmlw-retrieval-roberta-large-v2	100	27.53	47.49	51.97	32.33	36.07	39.08
stella-pl	100	23.20	45.82	52.34	27.58	41.45	38.08
stella-pl-retrieval-8k	100	23.23	43.30	48.40	28.17	34.00	35.42

Table 5: Evaluation results on clustering tasks using v-measure. The best score for a given column is marked in **bold**, and the second best is underlined.

Model name	Zero shot	SICK-E-PL	CDSC-E	PSC	PPC	Avg.
LaBSE	100	63.67	69.06	97.37	86.97	79.27
distiluse-base-multilingual-cased-v2	100	62.29	72.10	96.26	86.83	79.37
paraphrase-multilingual-MiniLM-L12-v2	100	71.78	72.39	97.07	92.37	83.40
paraphrase-multilingual-mpnet-base-v2	100	77.07	75.88	98.22	93.67	86.21
static-similarity-mrl-multilingual-v1	100	53.92	57.82	95.33	74.57	70.41
multilingual-e5-small	100	67.48	72.18	99.40	87.74	81.70
multilingual-e5-base	100	68.52	72.23	99.28	88.30	82.08
multilingual-e5-large	100	75.42	72.28	99.43	91.16	84.57
KaLM-embedding-multilingual-mini-instruct-v1	100	63.78	71.63	99.48	87.81	80.68
snowflake-arctic-embed-l-v2.0	100	63.24	71.02	99.48	87.08	80.20
snowflake-arctic-embed-m-v2.0	100	59.57	70.24	**99.54**	84.13	78.37
drama-base	100	54.05	60.18	95.53	78.45	72.05
drama-large	100	57.47	64.14	95.77	80.26	74.41
drama-1b	100	66.32	70.11	99.38	86.60	80.60
Qwen3-Embedding-0.6B	100	68.29	68.87	97.85	90.22	81.31
Qwen3-Embedding-4B	100	79.82	73.59	98.68	94.61	86.68
Qwen3-Embedding-8B	100	82.47	74.84	98.43	94.71	87.61
bge-multilingual-gemma2	100	**85.8**	78.51	99.27	**95.43**	**89.75**
silver-retriever-base-v1.1	100	55.84	62.67	98.75	82.04	74.82
st-polish-paraphrase-from-mpnet	100	80.39	75.17	99.03	93.67	87.06
st-polish-paraphrase-from-distilroberta	100	79.41	76.03	99.09	93.31	86.96
mmlw-e5-small	100	77.49	79.34	98.17	91.68	86.67
mmlw-e5-base	100	42.69	43.76	78.91	72.64	59.50
mmlw-e5-large	100	43.30	37.10	80.53	78.26	59.80
mmlw-roberta-base	100	81.85	79.23	98.59	92.97	88.16
mmlw-roberta-large	100	84.29	**79.96**	98.80	93.56	89.15
mmlw-retrieval-roberta-large	100	83.15	78.53	99.42	92.81	88.48
mmlw-retrieval-roberta-large-v2	100	79.27	75.61	**99.54**	91.69	86.53
stella-pl	100	84.68	79.20	99.31	93.60	89.20
stella-pl-retrieval-8k	100	85.66	79.26	**99.54**	93.77	89.56

Table 6: Evaluation results on pair classification tasks using average precision score based on cosine similarity. The best score for a given column is marked in **bold**, and the second best is underlined.

Model name	Zero shot	ArguAna-PL	DBPedia-PLHardNeg	FiQA-PL	HotpotQA-PLHardNeg	MSMARCO-PLHardNeg	NFCorpus-PL	NQ-PLHardNeg	Quora-PLHardNeg	SCIDOCS-PL	SciFact-PL	TRECCOVID-PL	Avg.
LaBSE	100	38.56	21.85	7.66	28.82	33.43	17.45	14.04	73.79	7.47	39.79	18.13	27.36
distiluse-base-multilingual-cased-v2	81	36.70	17.48	8.02	27.83	27.58	16.28	9.70	71.46	6.50	33.02	16.89	24.68
paraphrase-multilingual-MiniLM-L12-v2	81	37.86	22.34	12.49	28.86	38.43	17.17	15.95	76.61	10.26	40.23	34.22	30.40
paraphrase-multilingual-mpnet-base-v2	81	42.61	24.78	14.71	34.08	48.75	18.54	17.23	77.81	11.17	41.55	35.43	33.33
static-similarity-mrl-multilingual-v1	90	32.14	18.31	7.54	24.62	26.82	17.17	12.23	65.41	7.43	38.84	22.78	24.84
multilingual-e5-small	72	37.49	31.82	22.02	61.51	61.57	26.50	42.09	77.70	11.58	62.76	70.92	46.00
multilingual-e5-base	72	42.86	31.94	25.59	65.21	64.64	25.99	46.41	80.73	12.36	62.27	65.90	47.63
multilingual-e5-large	72	52.99	36.52	32.97	67.57	70.79	30.21	53.58	82.72	13.82	65.66	69.86	52.43
KaLM-embedding-multilingual-mini-instruct-v1	18	47.76	32.07	24.50	61.30	49.88	27.12	32.72	74.12	14.08	61.33	65.65	44.59
snowflake-arctic-embed-l-v2.0	81	54.61	39.73	36.85	66.58	69.58	32.11	52.13	83.59	17.04	67.94	76.98	54.29
snowflake-arctic-embed-m-v2.0	72	51.39	37.79	33.38	67.41	67.37	30.57	45.61	80.94	15.84	66.18	77.86	52.21
drama-base	72	40.58	8.13	11.49	29.35	21.29	21.35	3.65	64.35	11.22	58.05	41.73	28.29
drama-large	72	43.28	11.52	16.11	34.38	27.06	24.06	6.10	70.01	12.24	62.01	58.64	33.22
drama-1b	72	49.46	34.02	35.13	68.41	55.30	33.01	34.54	81.62	17.08	73.04	84.81	51.49
Qwen3-Embedding-0.6B	72	57.53	30.23	27.38	58.31	64.04	26.83	34.29	79.02	16.24	61.48	79.16	48.59
Qwen3-Embedding-4B	72	64.14	39.16	38.03	68.64	70.05	33.85	47.33	80.57	20.97	72.81	87.61	56.65
Qwen3-Embedding-8B	72	**66.82**	41.04	44.53	70.48	71.26	35.45	50.53	82.34	22.88	76.06	**89.93**	59.21
bge-multilingual-gemma2	54	59.24	43.67	45.44	**74.73**	74.00	36.89	57.42	84.08	18.08	73.45	81.26	58.93
silver-retriever-base-v1.1	100	47.07	31.69	24.99	49.85	62.15	29.29	42.34	78.40	11.04	52.80	42.53	42.92
st-polish-paraphrase-from-mpnet	100	51.86	29.13	22.28	36.27	50.35	24.04	26.12	80.61	13.24	52.47	35.22	38.33
st-polish-paraphrase-from-distilroberta	100	49.42	23.99	19.57	29.26	48.84	22.52	23.52	80.08	12.14	49.50	38.96	36.16
mmlw-e5-small	72	54.21	35.39	29.76	60.05	54.73	27.69	38.06	79.47	14.90	58.41	58.09	46.43
mmlw-e5-base	72	58.45	41.17	34.60	68.02	64.20	33.74	48.15	83.65	17.39	68.31	73.07	53.70
mmlw-e5-large	72	63.45	44.14	39.99	72.10	70.11	34.12	50.66	85.06	19.18	71.59	71.44	56.53
mmlw-roberta-base	90	59.04	40.33	35.21	68.30	64.07	34.17	49.25	83.79	17.95	66.00	71.48	53.60
mmlw-roberta-large	90	63.66	21.46	40.83	63.91	58.54	33.97	19.42	**86.05**	19.44	70.70	71.01	49.91
mmlw-retrieval-roberta-large	81	58.73	44.81	39.32	71.98	74.21	35.43	55.94	85.52	18.57	72.41	72.65	57.23
mmlw-retrieval-roberta-large-v2	54	61.04	43.34	44.91	68.99	71.47	37.48	59.81	82.05	21.60	74.63	76.49	58.35
stella-pl	54	60.22	**46.2**	**52.03**	71.18	72.62	**39.94**	61.62	85.67	**23.54**	78.3	77.72	60.82
stella-pl-retrieval-8k	54	66.03	44.36	51.51	73.84	**74.49**	39.56	**63.83**	85.07	22.57	**79.67**	76.54	**61.59**

Table 7: Evaluation results on retrieval tasks using nDCG@10. The best score for a given column is marked in **bold**, and the second best is underlined.

Model name	Zero shot	SICK-R-PL	CDSC-R	STSBenchmarkMultilingual	Avg.
LaBSE	100	65.90	85.53	72.58	74.67
distiluse-base-multilingual-cased-v2	100	65.53	87.67	74.06	75.75
paraphrase-multilingual-MiniLM-L12-v2	100	68.77	88.98	78.29	78.68
paraphrase-multilingual-mpnet-base-v2	100	73.13	88.80	81.46	81.13
static-similarity-mrl-multilingual-v1	100	61.40	86.97	67.65	72.01
multilingual-e5-small	100	70.62	90.95	73.67	78.41
multilingual-e5-base	100	71.46	89.61	76.32	79.13
multilingual-e5-large	100	74.86	89.80	79.57	81.41
KaLM-embedding-multilingual-mini-instruct-v1	100	66.58	90.00	72.13	76.24
snowflake-arctic-embed-l-v2.0	100	68.86	90.38	74.61	77.95
snowflake-arctic-embed-m-v2.0	100	66.57	90.22	70.00	75.60
drama-base	100	56.34	81.04	57.66	65.01
drama-large	100	58.76	83.39	58.99	67.05
drama-1b	100	69.81	89.72	75.09	78.21
Qwen3-Embedding-0.6B	100	69.63	88.32	77.40	78.45
Qwen3-Embedding-4B	100	77.85	91.43	<u>87.37</u>	85.55
Qwen3-Embedding-8B	100	80.11	91.60	**88.44**	<u>86.72</u>
bge-multilingual-gemma2	100	78.16	90.96	82.79	83.97
silver-retriever-base-v1.1	100	64.46	88.34	71.03	74.61
st-polish-paraphrase-from-mpnet	100	76.18	88.56	83.75	82.83
st-polish-paraphrase-from-distilroberta	100	76.37	89.62	81.89	82.63
mmlw-e5-small	100	74.66	90.57	80.91	82.05
mmlw-e5-base	100	43.11	59.57	44.39	49.02
mmlw-e5-large	100	33.98	40.00	45.86	39.95
mmlw-roberta-base	100	79.20	92.55	83.84	85.20
mmlw-roberta-large	100	79.91	92.54	83.25	85.23
mmlw-retrieval-roberta-large	100	79.36	**92.78**	82.00	84.71
mmlw-retrieval-roberta-large-v2	66	80.90	91.68	84.35	85.64
stella-pl	66	**81.92**	<u>92.68</u>	86.02	**86.87**
stella-pl-retrieval-8k	66	<u>81.65</u>	92.11	85.91	86.56

Table 8: Evaluation results on STS tasks using Spearman correlation based on cosine similarity. The best score for a given column is marked in **bold**, and the second best is <u>underlined</u>.