TR-MTEB: A Comprehensive Benchmark and Embedding Model Suite for Turkish Sentence Representations

Mehmet Selman Baysan
Boğaziçi University
Bebek, Istanbul
mehmet.baysan@std.bogazici.edu.tr

Tunga Güngör
Boğaziçi University
Bebek, Istanbul
gungort@bogazici.edu.tr

Abstract

We introduce TR-MTEB, the first large-scale, task-diverse benchmark designed to evaluate sentence embedding models for Turkish. Covering six core tasks as classification, clustering, pair classification, retrieval, bitext mining, and semantic textual similarity, TR-MTEB incorporates 26 high-quality datasets, including native and translated resources. To complement this benchmark, we construct a corpus of 34.2 million weakly supervised Turkish sentence pairs and train two Turkish-specific embedding models using contrastive pretraining and supervised fine-tuning. Evaluation results show that our models, despite being trained on limited resources, achieve competitive performance across most tasks and significantly improve upon baseline monolingual models. All datasets, models, and evaluation pipelines are publicly released[1] to facilitate further research in Turkish natural language processing and low-resource benchmarking.

1 Introduction

Text embeddings form the backbone of modern Natural Language Processing (NLP) systems, enabling efficient and scalable representations of text for a wide range of applications, from retrieval and classification to clustering and summarization. Recent advances have demonstrated that universal embedding models can power diverse downstream tasks with minimal adaptation, making them a critical component in the era of large-scale language technologies. However, the evaluation and development of such models have largely centered around high-resource languages like English, while low-resource, morphologically complex languages such as Turkish remain underrepresented.

Benchmarks like the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023) have played a pivotal role in standardizing the evaluation of text embeddings across multiple tasks and domains. Following the success of MTEB, language-specific extensions such as C-MTEB (Xiao et al., 2024) for Chinese and PL-MTEB (Poswiata et al., 2024) for Polish have shown that adapting benchmarks to linguistic and cultural contexts leads to deeper insights into model performance. Yet, no such comprehensive resource currently exists for Turkish, limiting both the development of robust Turkish embedding models and the fair evaluation of multilingual systems on Turkish text.

In this work, we introduce the Turkish Massive Text Embedding Benchmark (TR-MTEB), the first large-scale, standardized benchmark specifically designed to evaluate sentence embeddings for Turkish across six key tasks: classification, clustering, retrieval, bitext mining, pair classification, and semantic textual similarity. TR-MTEB draws from a diverse mixture of native Turkish datasets and high-quality translated resources, including carefully adapted retrieval datasets from the BEIR (Thakur et al., 2021) benchmark suite. All translations underwent rigorous quality control to ensure semantic fidelity and task alignment.

Beyond benchmarking, we address the critical shortage of high-quality Turkish training data for embedding models. Inspired by methodologies from the BGE (Xiao et al., 2024), GTE (Li et al., 2023) and E5 (Wang et al., 2022) models, we curated a corpus of 34.2 million weakly supervised Turkish sentence pairs sourced from publicly available data. Using this corpus, we pretrained general text embedding models for Turkish from BERTurk (Schweter, 2020) base model, and additionally trained these models using labeled data to see the effects of fine-tuning to assess their performance on TR-MTEB.

[1] TR-MTEB Github Repository

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 8867–8887
November 4-9, 2025

Our contributions are: (1) TR-MTEB, a task-diverse Turkish benchmark for evaluating sentence embeddings across six core NLP tasks; (2) a high-quality corpus of 34.2 million weakly supervised Turkish sentence pairs curated from over 50 public datasets; (3) newly trained Turkish-specific embedding models using contrastive pretraining and supervised fine-tuning on this corpus; and (4) publicly released datasets, models, and evaluation pipelines to support future Turkish NLP research and encourage equitable benchmarking for other low-resource languages.

2 Related Work

Recent years have seen the emergence of large-scale benchmarks like the Massive Text Embedding Benchmark (MTEB), which provides a unified framework for evaluating sentence embeddings across a wide range of NLP tasks, including classification, clustering, retrieval, and semantic similarity. Building on this foundation, language-specific adaptations such as C-MTEB for Chinese, PL-MTEB for Polish, MTEB-French (Ciancone et al., 2024), ruMTEB (Snegirev et al., 2024) for Russian, and Scandinavian Embedding Benchmark (Enevoldsen et al., 2024) have demonstrated the importance of tailoring benchmarks to linguistic nuances and resource availability.

However, for the Turkish language, efforts have been relatively limited. MUKAYESE (Safaya et al., 2022) has made progress by introducing task-specific benchmarks for Turkish NLP, including classification, translation, and summarization, but it does not evaluate sentence embeddings in a broad range of tasks. Additionally, earlier intrinsic evaluations of Turkish word embeddings (Arslan et al., 2023) highlighted the challenges posed by Turkish morphology, but they remained at the word level and lacked task diversity.

The BEIR benchmark (Thakur et al., 2021), widely used for zero-shot retrieval evaluation, has also inspired multilingual adaptations like BEIR-PL (Wojtasik et al., 2024) and BEIR-NL (Lotfi et al., 2025), which rely on translated datasets to support information retrieval in Polish and Dutch, respectively. These works show that translation, when combined with native resources, can effectively bootstrap evaluation frameworks for low-resource languages.

In addition to the development of benchmarks, recent advancements in embedding models have emphasized the importance of large-scale, weakly supervised training data. The C-Pack framework (Xiao et al., 2024) exemplifies this approach by introducing C-MTP, a massive dataset comprising 100 million Chinese text pairs extracted from diverse sources such as web corpora, question-answer (QA) forums, and encyclopedic content. These pairs were utilized to train the BAAI General Embeddings (BGE) models, which demonstrated significant performance improvements across various tasks in the C-MTEB benchmark.

Our work builds on these foundations by introducing TR-MTEB, the first task-diverse benchmark for evaluating sentence embeddings in Turkish. Like C-MTEB and PL-MTEB, we combine native corpora with carefully translated datasets, but we also go further by training and evaluating Turkish-specific embedding models, thus addressing a critical gap in Turkish NLP research.

Similar to C-MTP, our study involves the creation of 34.2 million weakly supervised Turkish sentence pairs from various sources. These pairs were instrumental in training Turkish-specific embedding models, with the aim of improving performance across the tasks defined in the TR-MTEB benchmark. This approach aligns with the strategies employed in the C-Pack and E5 frameworks, underscoring the efficacy of large-scale, diverse training data in developing versatile embedding models.

In parallel to these efforts, the Massive Multilingual Text Embedding Benchmark (MMTEB) has recently been introduced as a large-scale, community-driven expansion of MTEB (Enevoldsen et al., 2025). Covering more than 500 tasks across over 250 languages, MMTEB broadens the scope of embedding evaluation by incorporating novel and challenging tasks such as instruction following, long-document retrieval, and code retrieval. To mitigate the computational cost of such a massive benchmark, the authors propose correlation-based downsampling and optimized retrieval splits with hard negative sampling, while still preserving model ranking fidelity. This concurrent work highlights the growing importance of multilingual and large-scale benchmarking for embeddings. In contrast, our TR-MTEB benchmark focuses specifically on Turkish, addressing linguistic challenges such as rich morphology and limited resources, thereby complementing the multilingual perspective of MMTEB with a fine-grained, language-specific evaluation framework.

3 Methodology

In this section, we present our methodology for developing and evaluating Turkish sentence embedding models. Our approach consists of three main components: (1) the construction of a comprehensive Turkish benchmark for embedding evaluation, (2) the collection and curation of a large-scale weakly supervised text pair corpus for model training, and (3) the training of Turkish-specific embedding models using modern contrastive learning techniques. Together, these components aim to establish a robust evaluation and training framework tailored to the linguistic characteristics and resource limitations of Turkish.

3.1 Turkish Massive Text Embedding Benchmark (TR-MTEB)

The motivation for constructing the Turkish Massive Text Embedding Benchmark (TR-MTEB) stems from the limited coverage of Turkish within existing multilingual embedding evaluation frameworks. At the time this work was initiated, the multilingual version of the MTEB[2] benchmark included only two datasets for classification and a single dataset for retrieval in Turkish. This limited scope posed a serious challenge: evaluating the performance of a Turkish-specific sentence embedding model using only these datasets would provide an incomplete and potentially misleading picture of its true capabilities. Conversely, evaluating such a model on datasets designed for other languages would fail to capture the unique syntactic, morphological, and semantic features of Turkish.

To address this gap, TR-MTEB was developed as a task-diverse and language-specific benchmark that assembles high-quality datasets tailored for Turkish. We curated publicly available datasets from prior academic studies, selecting only those with associated peer-reviewed publications. In addition to this, our selection process also considered data quality and cleanliness (favoring datasets with minimal preprocessing requirements and consistent schema), size and diversity (including both large-scale and smaller domain-specific datasets), licensing and accessibility (restricting to openly licensed resources to ensure reproducibility), and citation and community usage (preferring datasets with greater academic adoption and reliability). TR-MTEB includes evaluation tasks across six core NLP applications: classification, clustering, pair classification, retrieval, bitext mining, and semantic textual similarity (STS). Each task is adapted from the evaluation design of the original MTEB benchmark and restructured where needed to align with Turkish language characteristics.

A brief description of how each task is formulated in TR-MTEB is provided below. An overview of all task categories and their associated datasets is illustrated in Figure 1.

- **Classification:** Sentence embeddings from the train set are used to train a logistic regression classifier, which is then evaluated on the test set. The primary metric is accuracy, along with average precision and F1-score.

- **Clustering:** This task groups semantically similar sentences or paragraphs. Sentence embeddings are clustered using a mini-batch k-means algorithm, with the number of clusters equal to the number of unique labels. The primary metric is V-measure, which balances homogeneity and completeness.

- **Pair Classification:** For tasks such as paraphrasing or duplicate detection, the cosine similarity between paired embeddings is computed and thresholded. The primary metric is average precision, with additional metrics including accuracy, F1-score, precision, and recall.

- **Bitext Mining:** The goal is to identify translation pairs across two languages. Sentence embeddings are generated for each side of the corpus and cosine similarity is used to find the most likely translations. The primary metric is F1-score, additional metrics precision, recall, and accuracy are also reported.

- **Semantic Textual Similarity (STS):** Given a sentence pair, the task is to measure their semantic similarity on a continuous scale. Cosine similarity between embeddings is correlated with gold-standard similarity scores. The primary metric is Spearman correlation, supplemented by Pearson correlation.

- **Retrieval:** This task ranks a document corpus for each query based on embedding similarity. The primary metric is nDCG@10, additionally MRR@k, MAP@k, precision@k, and recall@k are calculated.

[2] Multilingual MTEB Leaderboard

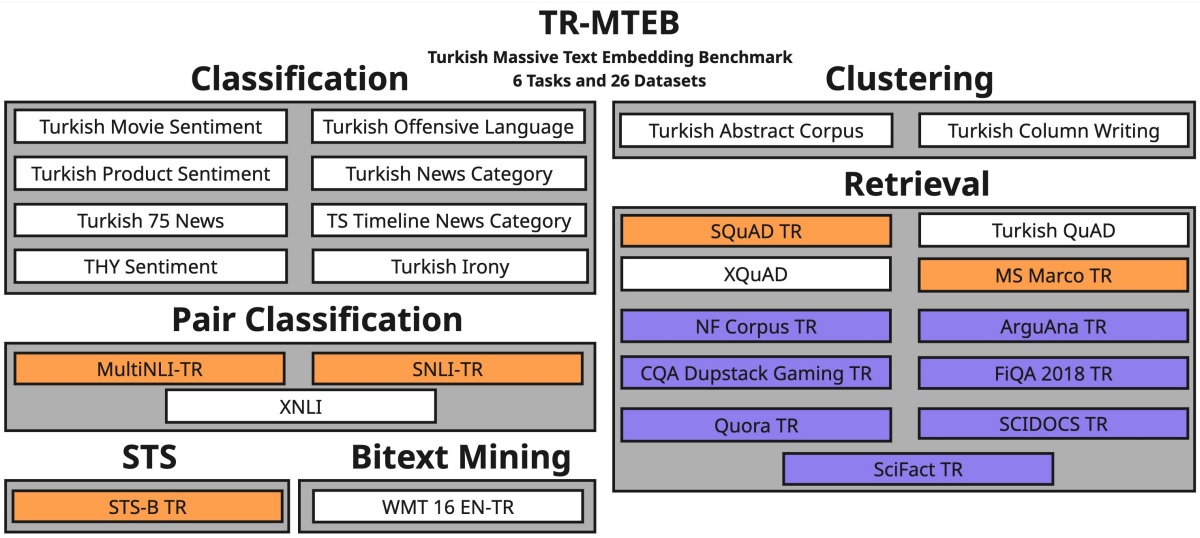

Figure 1: An overview of tasks and datasets included in TR-MTEB. Datasets in white background are original Turkish datasets. Datasets highlighted in orange indicate those that were translated to Turkish prior to this study, while datasets marked in violet represent English datasets that were translated into Turkish as part of this work. A detailed description and key statistics of all datasets included in the benchmark can be found in Appendix A.

By expanding the range of tasks and datasets for Turkish, TR-MTEB provides a reliable and comprehensive framework for evaluating embedding models tailored to or tested on the Turkish language. In the following subsections, we describe how the benchmark was assembled, including both the collection of existing Turkish datasets and the translation of key retrieval datasets from the BEIR benchmark.

3.1.1 Collecting Existing Datasets

To build TR-MTEB, we curated publicly available Turkish datasets spanning six core NLP tasks. We prioritized resources validated through peer-reviewed publications to ensure annotation quality and linguistic diversity reflective of real-world applications.

For **classification**, we selected six datasets covering the topics of news categorization (Amasyali and Yildirim, 2004; Sak et al., 2008), sentiment analysis (Sezer, 2022), offensive speech detection (Çöltekin, 2020), and irony detection (Karagoz et al., 2021). Two classification datasets from the multilingual MTEB benchmark were also included for compatibility. The **clustering** task is supported by two datasets (Amasyalı and Diri, 2006; Özturk et al., 2014), each composed of Turkish news articles and opinion pieces. For **pair classification**, we included machine-translated versions of MNLI and SNLI (Budur et al., 2020) and the Turkish split of XNLI (Conneau et al., 2018), using binary labels to represent entailment vs. non-entailment. **Bitext mining** is evaluated using the English–Turkish subset of WMT-16 (Bojar et al., 2016), enabling alignment-based evaluation of cross-lingual similarity. The **STS** task uses a translated version of STS-Benchmark (Beken Fikri et al., 2021), providing graded similarity scores for Turkish sentence pairs. Lastly, the **retrieval** task is supported by four datasets which are Turkish-translated versions of SQuAD (Budur et al., 2024) and MS MARCO (Kesgin et al., 2023), a domain-specific Turkish QA dataset focused on historical topics (Soygazi et al., 2021), and the Turkish split of the XQuAD benchmark. These datasets enable the evaluation of retrieval performance across general and domain-specific settings.

3.1.2 Translation of Retrieval Datasets from BEIR Benchmark

To expand retrieval coverage in TR-MTEB, we translated seven diverse datasets from the BEIR benchmark into Turkish, covering domains like question answering, fact verification, and web search. This approach follows trends in recent multilingual benchmarks (e.g., PL-MTEB), which incorporate high-quality machine translations to support low-resource languages.

Given the recent advances in large language models (LLMs) and their superior performance over traditional translation methods in downstream NLP tasks, we opted to employ an LLM-based transla-

tion strategy. To determine the most suitable model for Turkish, we relied on the Cetvel Benchmark (Kesen et al., 2024), a unified evaluation suite designed to benchmark LLMs specifically for the Turkish language.

In the machine translation category of Cetvel, the 35B parameter version of Cohere's aya-expanse model (Dang et al., 2024) achieved the highest score. However, due to computational constraints, we selected the 8B version of the same model, which achieved nearly equivalent translation quality while offering significantly better runtime efficiency. Using this model, we translated both the query sets and the corpora of the selected BEIR datasets into Turkish, following the structured translation protocol detailed in Appendix B.1.

To ensure translation quality, we developed an LLM-as-a-Judge evaluation pipeline designed to automatically assess the semantic adequacy of machine-translated datasets. To calibrate this evaluation setup, we first sampled a total of 115 examples, balanced across all translated datasets and including both queries and corpus entries. Each example was manually annotated by one primary annotator, with spot checks by a second reviewer, and labeled as either `PASS` or `FAIL` according to predefined guidelines focusing on semantic fidelity (whether the meaning was accurately preserved) and fluency (grammaticality and naturalness in Turkish). The annotation covered 115 sentence pairs evenly drawn from seven datasets and required approximately six hours in total, conducted in randomized batches to minimize bias. We then used these labeled samples to iteratively refine the evaluation prompt issued to the LLM, aiming to align its judgments with human annotations. The finalized prompt, when applied to the same 115 examples, achieved 85.2% agreement with human labels, with a precision of 92.9%, recall of 84.4%, and an F1 score of 88.4% (see Figure 2 and Appendix B.2).

Upon validation, this evaluator was applied at scale. Samples marked as FAIL were manually corrected or re-translated using larger models from Google AI Studio[3]. This hybrid approach ensured accurate and fluent translations, allowing us to robustly extend the retrieval task in TR-MTEB with high-quality Turkish datasets.

In total, TR-MTEB comprises **26 datasets** spanning **6 task** categories: 8 datasets for classification, 2 for clustering, 3 for pair classification, 1 for bitext mining, 1 for STS, and 11 for retrieval. This comprehensive collection ensures that Turkish embedding models can be evaluated across a broad range of semantic understanding and downstream utility tasks.

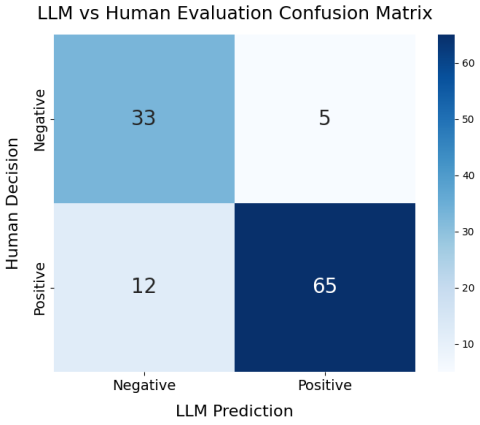

Figure 2: Confusion matrix comparing human annotations and LLM predictions for translation quality evaluation.

3.2 Large-Scale Turkish Sentence Pair Corpus

During the initial stages of our benchmarking efforts, we identified a significant gap in the availability of Turkish-specific sentence embedding models. While multilingual models are widely available, there existed no large-scale, publicly trained embedding model tailored specifically for Turkish. To address this shortcoming, both for the purpose of training embedding models within this study and to support future research in the community, we constructed a large-scale high-quality corpus of 34.2 million Turkish weakly supervised sentence pairs from publicly accessible sources.

Our primary objective was to collect semantically aligned sentence pairs spanning diverse domains such as general discourse, technology, healthcare, news, and education. This diversity is crucial to ensure that the resulting embeddings generalize beyond narrow domains and perform robustly across a wide range of tasks. We sourced the data primarily from the Hugging Face Datasets Hub[4], filtering for Turkish-language datasets that are suitable for sentence pair construction. These include original Turkish datasets, machine-translated corpora, and synthetically gen-

[3]Google AI Studio

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

erated resources using LLMs. A detailed list of the dataset sources, their schema types, and filtering statistics are provided in Appendix D.

Since the datasets varied in structure and format, we applied a set of heuristics for constructing sentence pairs based on their schema:

- Question-Answer datasets: When datasets consisted of question-answer pairs, we directly used the (question, answer) pairs and additionally created (question, context) and (answer, context) pairs if contextual information was available as sentence pairs.

- Title-Content datasets: If a dataset included a title and a corresponding content, we used these as sentence pairs. In cases where a summary was also provided, we generated combinations such as (summary, content) and (title, summary) to ensure both symmetric and asymmetric semantic alignment.

- Paraphrase or entailment datasets: For resources annotated with semantically equivalent or related sentences, we extracted all sentence pairs that conveyed the same meaning, regardless of directionality.

The resulting dataset includes a deliberate mix of symmetric and asymmetric sentence pairs to reflect the diversity of real-world language understanding tasks. Symmetric pairs represent bidirectional semantic equivalence, while asymmetric pairs capture directional relationships such as query-passage or question-answer structures. This design ensures that models are exposed to a broad spectrum of sentence relationships, supporting robust training for downstream applications including retrieval, STS, and pair classification.

We applied a multi-stage filtering process to the collected 62.5 million sentence pairs to improve the quality of the constructed corpus while preserving its scale. Initially, we drew equal-sized random subsets from each source dataset and evaluated pairwise semantic similarity using the top-performing multilingual embedding models E5, BGE, and GTE from the MTEB leaderboard. However, the similarity scores of the semantically related and unrelated sentence pairs were often indistinguishably close, making it difficult to establish a meaningful threshold for effective filtering.

Based on this observation, to obtain a suitable model for filtering, we first trained three candidate sentence embedding models using approximately 2–3 million high-quality Turkish sentence pairs drawn from three academically validated datasets (Scialom et al., 2020), (Özturk et al., 2014), (Baykara and Güngör, 2022). These models, based on `multilingual-e5-base` and cased and uncased versions of BERTurk, were contrastively pre-trained using Multiple Negatives Ranking Loss (MNRL) with a batch size of 64. Among these candidates, the contrastively trained `multilingual-e5-base` model exhibited the most consistent and strongest performance on the proposed TR-MTEB benchmark (see Table 1) where *Mean (Task)* refers to the average score across datasets and *Mean (Task Type)* is averaged over task categories.

Model	Mean (Task)	Mean (Task Type)
`multilingual-e5-base contrastive_pretrained`	57.77	65.92
`bert-base-turkish-uncased contrastive_pretrained`	53.62	53.38
`bert-base-turkish-cased contrastive_pretrained`	53.73	53.43
`multilingual-e5-base fine_tuned`	62.73	67.96

Table 1: Performance of small-scale trained Turkish embedding models on the TR-MTEB benchmark. These models were evaluated only for internal comparison and not for downstream use. For detailed benchmarking of final models, see Table 2.

To improve filtering effectiveness, we further fine-tuned the selected `multilingual-e5-base` model on the supervised datasets. This fine-tuned model achieved 62.73 Mean (Task) and 67.96 Mean (Task Type) on TR-MTEB, outperforming its contrastively pre-trained version by a clear margin. Using this model, we re-evaluated representative samples from the larger corpus. Following a manual inspection of the resulting similarity distributions, we empirically determined a similarity threshold of 0.4 as the most inclusive yet reasonably precise cutoff. While some semantically relevant pairs with low surface similarity may have been excluded, the threshold of 0.4 provided the best trade-off between coverage and noise reduction.

We then applied this fine-tuned model to compute similarity scores over the entire 62.5 million sentence pairs corpus[5] and discarded all pairs with similarity scores below the 0.4 threshold. This multi-stage filtering procedure significantly im-

[5] Full Corpus HuggingFace Link

proved the overall quality of the corpus, yielding a final dataset of 34.2 million high-quality sentence pairs[6] that span diverse domains. The resulting corpus serves as a robust foundation for contrastive pretraining and is well-suited for downstream tasks such as semantic similarity, retrieval, and classification in Turkish.

The final corpus represents the largest publicly documented collection of Turkish sentence pairs to date and is made available to support further research on Turkish embedding models, contrastive learning approaches, and low-resource NLP development.

3.3 Training Turkish Sentence Embedding Models

To assess the value of our sentence pairs corpus and establish robust TR-MTEB baselines, we trained a Turkish embedding model initialized from BERTurk[7], a monolingual BERT model tailored for Turkish. In contrast to multilingual encoders requiring large-scale compute, our objective was to test whether high-quality Turkish-only training could yield competitive results.

The model was pretrained on the full 34.2M filtered sentence pairs using Cached Multiple Negatives Ranking Loss (Cached-MNRL), which simulates large batch contrastive learning by caching embeddings across iterations (Gao et al., 2021). This allowed us to reach an effective batch size of 32,768 while maintaining efficiency on single GPU. A detailed overview of hyperparameters, training duration, and compute resources for both the contrastive pretraining and supervised fine-tuning stages is provided in Appendix E.

The loss is defined as:

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\exp\left(\cos(a_i, p_i)/\tau\right)}{\sum_{j=1}^{n}\exp\left(\cos(a_i, p_j)/\tau\right)} \quad (1)$$

where a_i and p_i are embeddings of anchor and positive sentences, respectively, and τ is a temperature hyperparameter. The denominator spans all positive samples in the batch, treating p_j with $j \neq i$ as implicit negatives. By leveraging in-batch negatives, MNRL enables efficient and effective contrastive learning without the need for explicit negative sampling.

Following contrastive pretraining, we fine-tuned the model on task-specific Turkish datasets to better adapt the model to task-specific requirements spanning classification, pair classification, STS, and retrieval, using TR-MTEB training sets corresponding to MS MARCO, STS-B, SNLI, and WMT-16 tasks. Depending on the task, we used appropriate objectives: MNRL for retrieval, softmax cross-entropy for classification, and CoSENT loss for similarity scoring.

A modular evaluation pipeline was employed, ensuring fair metric alignment per task. This combined approach of unsupervised pretraining and supervised fine-tuning yielded models well-suited to the linguistic and structural nuances of Turkish, establishes strong baselines for TR-MTEB, and are publicly available to support further research.

4 Experiments & Results

To evaluate sentence embedding models for Turkish, we benchmarked a diverse set of multilingual and monolingual encoders on the TR-MTEB benchmark. Table 2 reports the results across the six core tasks: Bitext Mining, Classification, Clustering, Pair Classification, Retrieval, and Semantic Textual Similarity (STS). The evaluated models include publicly available state-of-the-art encoders, static word embedding baselines, and our in-house Turkish-specific models. Evaluations were conducted using a Turkish-adapted version of the official MTEB library, ensuring compatibility with original metrics while enabling standardized assessment on TR-MTEB.

Model Selection Strategy: We include a diverse range of models to capture differences in architecture, training objective, and language coverage (see Appendix C for details). Multilingual SOTA models like `multilingual-e5` and `gte` help us assess high-capacity general-purpose encoders. Instruction-tuned variants (e.g., `e5-large-instruct`) are included to evaluate the benefits of task-specific supervision. SBERT-based models such as `LaBSE` and `mpnet/MiniLM` represent widely adopted lightweight alternatives. We also test monolingual Turkish models like `bert-base-turkish-uncased` to establish native-language baselines and analyze their extendability. Static word embeddings (FastText, GloVe) offer insight into performance gaps between contextual and traditional approaches. Lastly, we include `text-embedding-3-small` to compare against a proprietary embedding model.

[6] Filtered Corpus HuggingFace Link
[7] BERTurk Model

Model	Dim	Mean (Task)	Mean (Type)	Bitext	Class.	Clust.	Pair Cl.	Retr.	STS
multilingual-e5-large	1024	66.82	73.07	99.43	71.77	60.42	65.01	60.62	81.18
multilingual-e5-large-instruct	1024	65.92	72.78	98.99	73.70	61.78	63.74	57.21	81.23
gte-multilingual-base	1024	64.49	71.96	97.98	67.79	60.39	67.41	57.51	80.68
multilingual-e5-base	768	64.26	70.97	98.51	68.20	60.25	62.20	58.29	78.37
text-embedding-3-small	768	66.61	69.52	91.55	69.51	62.08	58.13	64.99	70.85
turkish-embedding-model-fine-tuned	768	63.42	70.25	86.88	70.93	61.49	71.03	52.75	78.44
paraphrase-multilingual-mpnet-base-v2	768	61.71	71.05	96.06	69.06	57.52	72.22	49.27	82.18
multilingual-e5-small	384	62.53	69.55	97.46	65.64	59.14	62.34	56.53	76.18
turkish-embedding-model	768	59.84	59.49	37.7	70.68	63.45	57.65	52.54	74.93
LaBSE	768	57.60	66.60	99.53	65.60	58.40	57.83	46.47	71.75
paraphrase-multilingual-MiniLM-L12-v2	768	58.94	68.44	93.67	65.76	56.64	69.56	46.57	78.41
bert-base-turkish-uncased	768	49.54	47.07	7.07	70.12	60.77	52.92	34.82	56.74
fasttext-turkish	300	46.93	44.18	4.52	63.35	59.74	52.70	34.67	50.08
glove-turkish	-	41.37	39.82	0.22	62.54	58.41	49.48	24.14	44.12
all-MiniLM-L12-v2	384	38.65	38.50	8.27	50.63	39.95	51.79	27.60	52.78
all-MiniLM-L6-v2	384	37.06	37.02	6.78	52.46	40.57	50.52	23.25	48.55
all-mpnet-base-v2	768	37.62	38.06	5.07	54.43	46.63	49.79	22.26	50.21

Table 2: TR-MTEB benchmark results across embedding models. The row highlighted in light gray indicates the models trained in this study. Bitext: Bitext Mining, Class.: Classification, Clust.: Clustering, Pair Cl.: Pair Classification, Retr.: Retrieval, STS: Semantic Textual Similarity. Results ordered by Mean (Task) scores.

Our Models: We introduce two Turkish-specific models trained from scratch using the `bert-base-turkish-uncased` backbone. All models were implemented using the Python programming language and `transformers` (Wolf et al., 2020), `datasets` (Lhoest et al., 2021), and `sentence-transformers` (Reimers and Gurevych, 2019) libraries:

1. **`turkish-embedding-model`**[8]: Contrastively trained on 34.2M filtered sentence pairs using Cached-MNRL with a batch size of 32,768. Training was performed on a single A100 GPU over approximately 80 GPU hours via Google Colab[9].

2. **`turkish-embedding-model-fine-tuned`**[10]: Further fine-tuned using labeled datasets from TR-MTEB to adapt to task-specific objectives. Fine-tuning took approximately 2 GPU hours using the same A100 setup.

Both models are highlighted in light gray in Table 2. The fine-tuned version achieves a notable improvement across tasks, increasing Mean(Task) from 59.84 to 63.42 and Mean(Type) from 59.49 to 70.25.

Effect of Domain-Specific Training: Compared to their base model `bert-base-turkish-uncased`, which scores 49.54 Mean(Task) and 47.07 Mean(Type), our models demonstrate significant improvements across all task types. This validates the effectiveness of large-scale contrastive training with high-quality Turkish sentence pairs in enhancing general-purpose semantic representations.

Bitext Mining Challenge: One notable performance gap appears in the Bitext Mining task, where our models score 37.7 and 86.88, respectively, substantially below multilingual counterparts such as `e5-large` (99.43) or `LaBSE` (99.53). This is expected, as our models are monolingual and lack exposure to English or other cross-lingual sentence pairs during pretraining. As a result, they underperform in aligning Turkish–English sentence representations, which is crucial for Bitext Mining.

Impact of Fine-Tuning: Supervised fine-tuning significantly boosts downstream performance. The fine-tuned model shows large gains in Bitext Mining (+49.2 points), Pair Classification (+13.4 points), and STS (+3.5 points) over its contrastive-only version. This highlights the benefit of task-aware adaptation even when the underlying encoder is trained on large-scale unsupervised data.

Comparison with Multilingual Models: While the best performing multilingual models (`e5-large`, `gte`) still lead in overall performance, benefiting from web-scale multilingual corpora and up to 560M parameters, our Turkish-specific model of 110M parameter significantly narrows the gap, especially in tasks where deep linguistic

[8] Turkish Embedding Model HuggingFace Link
[9] Google Colab
[10] Fine-Tuned Turkish Embedding Model HuggingFace Link

alignment is more important than cross-lingual generalization (e.g., classification, clustering, and STS).

Our findings show that compact, monolingual encoders trained on linguistically and semantically rich Turkish corpora can yield competitive results across a wide spectrum of NLP tasks. While further improvements in Bitext and zero-shot generalization would require cross-lingual training, our models already provide strong baselines for Turkish NLP and demonstrate the viability of low-resource, high-quality training pipelines.

5 Conclusion

This paper introduced TR-MTEB, the first large-scale benchmark for evaluating Turkish sentence embeddings across six task types and 26 datasets, combining native and high-quality translated resources. To support training, we compiled a 34.2M weakly supervised sentence pairs corpus and used it to train two Turkish-specific embedding models based on `bert-base-turkish-uncased`: one trained with Cached-MNRL for efficient contrastive learning, and one further fine-tuned on TR-MTEB training splits.

Our models substantially outperform the base BERTurk model and perform competitively on monolingual tasks such as classification and STS, despite being trained with significantly less data and compute than large multilingual counterparts. The remaining gap in cross-lingual tasks, notably bitext mining, highlights the limitations of monolingual-only training for multilingual alignment.

As future work, we plan to focus on scaling the training data, incorporating multilingual contrastive objectives to improve cross-lingual generalization, and extending TR-MTEB with additional tasks such as reranking and summarization to enable more comprehensive evaluation.

6 Limitations

While TR-MTEB represents a significant step forward for benchmarking sentence embeddings in Turkish, our work has several limitations that future research should address.

First, a substantial portion of TR-MTEB is composed of machine-translated datasets, despite manual inspection and LLM-based validation (Section 3.1.2), datasets may still differ from native Turkish in terms of syntactic structure, discourse coherence, or idiomatic expressions. While this approach enables broader task coverage, especially for retrieval and entailment tasks with no native Turkish equivalents, it may introduce subtle artifacts. It is worth noting that this reliance on high-quality translation is not unique to Turkish: several language-specific BEIR extensions, such as BEIR-PL and BEIR-NL, also use translated benchmarks to compensate for dataset scarcity in low-resource languages.

Second, the sentence pairs corpus used for model training, while large in scale (34.2M pairs), is constructed from weak supervision. Despite filtering via embedding-based similarity scoring, the resulting dataset may still contain noise or spurious patterns that limit the upper bound of model performance. Moreover, while recent high-resource language studies (e.g., English or Chinese) often rely on hundreds of millions to billions of pairs, our Turkish corpus remains relatively small in comparison. This limitation stems from the scarcity of large-scale public resources in Turkish, but we believe that future efforts to expand this corpus, both in scale and domain coverage, could further improve embedding quality. Additionally, our models were trained only on Turkish text, which restricts their ability to perform well on inherently cross-lingual tasks such as bitext mining.

Third, our embedding models were trained from a relatively lightweight Turkish monolingual encoder (`bert-base-turkish-uncased`), which limits their parameter capacity compared to multilingual state-of-the-art models such as `e5-large` or `GTE`. While our findings demonstrate that domain-specific training can bridge part of this gap, performance on certain tasks, especially STS and bitext mining, remains behind larger multilingual models.

Lastly, TR-MTEB does not yet include tasks such as summarization, re-ranking, or question generation, which are important for real-world applications. We also do not currently assess the robustness of models to noisy input or domain shifts.

We hope that future extensions of TR-MTEB will expand the task coverage, include more native Turkish datasets, and enable broader comparison across model architectures, training paradigms, and languages.

References

Duygu Altinok. 2024. Instructurca: A diverse instructional content dataset for turkish.

M. Fatih Amasyalı and Banu Diri. 2006. Automatic turkish text categorization in terms of author, genre and gender. In *Natural Language Processing and Information Systems*, pages 221–226, Berlin, Heidelberg. Springer Berlin Heidelberg.

M.F. Amasyali and T. Yildirim. 2004. Automatic text categorization of news articles. In *Proceedings of the IEEE 12th Signal Processing and Communications Applications Conference, 2004.*, pages 224–226.

MF Amasyalı and T Yıldırım. 2004. Otomatik haber metinleri sınıflandırma. *SIU 2004*, pages 224–226.

Oğuz Ali Arslan, Berfin Duman, Hakan Erdem, Can Günyel, Bike Sönmez, and Doğukan Arslan. 2023. Towards turkish word embeddings: An intrinsic evaluation. *2023 8th International Conference on Computer Science and Engineering (UBMK)*, pages 564–568.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Batuhan Baykara and Tunga Güngör. 2022. Abstractive text summarization and new large-scale datasets for agglutinative languages turkish and hungarian. *Lang. Resour. Eval.*, 56(3):973–1007.

Ali Bayram. 2024a. Onedio haberler veri seti.

M. Ali Bayram. 2024b. Turkish MMLU: Yapay Zeka ve Akademik Uygulamalar İçin En Kapsamlı ve Özgün Türkçe Veri Seti.

M. Ali Bayram. 2024c. Türkçe Tıbbi Soru-Cevap Veri Seti: 167 Bin Sağlık Sorusu ve Cevabı.

Figen Beken Fikri, Kemal Oflazer, and Berrin Yanikoglu. 2021. Semantic similarity based evaluation for abstractive news summarization. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 24–33, Online. Association for Computational Linguistics.

Abdullah Bezir. 2024. bezir/gsm8k-tr. https://huggingface.co/datasets/bezir/bezir/gsm8k-tr.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Ond rej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, and 2 others. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval.

Emrah Budur, Rıza Özçelik, Dilara Soylu, Omar Khattab, Tunga Güngör, and Christopher Potts. 2024. Building efficient and effective openqa systems for low-resource languages. *Know.-Based Syst.*, 302(C).

Emrah Budur, Rıza Özçelik, Tunga Gungor, and Christopher Potts. 2020. Data and Representation for Turkish Natural Language Inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8253–8267, Online. Association for Computational Linguistics.

Muhammed Kayra Bulut. 2024. Patient doctor q&a tr 19583.

Çağrı Çöltekin. 2020. A corpus of turkish offensive language on social media. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France.

Mathieu Ciancone, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. 2024. Extending the massive text embedding benchmark to french. *ArXiv*, abs/2405.20468.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur P. Parikh. 2023. Seahorse: A multilingual, multifaceted dataset for summarization evaluation. *Preprint*, arXiv:2305.13194.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Lili Jiang Meg Risdal Nikhil Dandekar tomtung DataCanary, hilfialkaff. 2017. Quora question pairs.

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2021. MFAQ: a multilingual FAQ dataset. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 1–13, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Erkin Demirtas and Mykola Pechenizkiy. 2013. Cross-lingual polarity detection with machine translation. In *wisdom*.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, and 67 others. 2025. Mmteb: Massive multilingual text embedding benchmark. *Preprint*, arXiv:2502.13595.

Kenneth Enevoldsen, Márton Kardos, Niklas Muennighoff, and Kristoffer Laigaard Nielbo. 2024. The scandinavian embedding benchmarks: Comprehensive assessment of multilingual and monolingual text embedding. In *Advances in Neural Information Processing Systems*, volume 37, pages 40336–40358. Curran Associates, Inc.

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.

Wikimedia Foundation. Wikimedia downloads.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. 2021. Scaling deep contrastive learning batch size under memory limited setup. *Preprint*, arXiv:2101.06983.

Musab Gultekin. 2023. Wikipedia turkish summarization dataset. https://huggingface.co/datasets/musabg/wikipedia-tr-summarization.

Srishti Gureja, Lester James V Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. 2024. M-rewardbench: Evaluating reward models in multilingual settings. *arXiv preprint arXiv:2410.15522*.

Doris Hoogeveen, Karin M. Verspoor, and Timothy Baldwin. 2015. Cqadupstack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian Document Computing Symposium (ADCS)*, ADCS '15, pages 3:1–3:8, New York, NY, USA. ACM.

Andreea Iana, Goran Glavaš, and Heiko Paulheim. 2024. Mind your language: a multilingual dataset for cross-lingual news recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 553–563.

Pinar Karagoz, Asli Umay Ozturk, and Yesim Cemek. 2021. Ironytr: Irony detection in turkish informal texts. *Int. J. Intell. Inf. Technol.*, 17(4):1–18.

Ilker Kesen, Mustafa Cemil Guney, Aykut Erdem, and Gozde Gul Sahin. 2024. Cetvel: A unified benchmark for evaluating turkish llms.

Himmet Toprak Kesgin, Muzaffer Kaan Yuce, and Mehmet Fatih Amasyali. 2023. Developing and evaluating tiny to medium-sized turkish bert models. *arXiv preprint arXiv:2307.14134*.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

Felix Leeb and Bernhard Schölkopf. 2024. A diverse multilingual news headlines dataset from around the world. *Preprint*, arXiv:2403.19352.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, and 13 others. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *ArXiv*, abs/2308.03281.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Openorca: An open dataset of gpt augmented flan reasoning traces. https://https://huggingface.co/datasets/Open-Orca/OpenOrca.

Ehsan Lotfi, Nikolay Banar, and Walter Daelemans. 2025. BEIR-NL: Zero-shot information retrieval benchmark for the Dutch language. In *Proceedings of the 18th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 36–45, Abu Dhabi, UAE. 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.

OpenAI. 2024. New embedding models and api updates. Accessed: 2025-05-14.

Seçil Özturk, Bülent Sankur, Tunga Gungör, Mustafa Berkay Yilmaz, Bilge Köroğlu, Onur Ağin, Mustafa İşbilen, Çağdaş Ulaş, and Mehmet Ahat. 2014. Turkish labeled text corpus. In *2014 22nd Signal Processing and Communications Applications Conference (SIU)*, pages 1395–1398. IEEE.

Chester Palen-Michel and Constantine Lignos. 2023. LR-sum: Summarization for less-resourced languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6829–6844, Toronto, Canada. Association for Computational Linguistics.

Rafal Poswiata, Slawomir Dadas, and Michal Perelkiewicz. 2024. Pl-mteb: Polish massive text embedding benchmark. *ArXiv*, abs/2405.10138.

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.

Ali Safaya, Emirhan Kurtuluş, Arda Goktogan, and Deniz Yuret. 2022. Mukayese: Turkish NLP strikes back. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 846–863, Dublin, Ireland. Association for Computational Linguistics.

Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2008. Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *Advances in Natural Language Processing*, pages 417–427, Berlin, Heidelberg. Springer Berlin Heidelberg.

Karahan Sarıtaş, Cahid Arda Öz, and Tunga Güngör. 2024. A comprehensive analysis of static word embeddings for turkish. *Expert Systems with Applications*, 252:124123.

Yves Scherrer. 2020. TaPaCo: A Corpus of Sentential Paraphrases for 73 Languages.

Stefan Schweter. 2020. Berturk - bert models for turkish.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

sekerlipencere. 2024. zynpdata: Türkiye'nin en büyük açık kaynaklı türkçe forum veri seti. https://github.com/sekerlipencere/zynpdata-zynp_ai-teknofest.

Taner Sezer. 2022. Thy-sa dataset.

Artem Snegirev, Maria Tikhonova, Anna Maksimova, Alena Fenogenova, and Alexander Abramov. 2024. The russian-focused embedders' exploration: rumteb benchmark and russian embedding model design. *ArXiv*, abs/2408.12503.

Fatih Soygazi, Okan Çiftçi, Uğurcan Kök, and Soner Cengiz. 2021. Thquad: Turkish historic question answering dataset for reading comprehension. In *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pages 215–220.

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 *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Mesut Erhan Unal, Begum Citamak, Semih Yagcioglu, Aykut Erdem, Erkut Erdem, Nazli Ikizler Cinbis, and Ruket Cakici. Tasviret: A benchmark dataset for automatic turkish description generation from images. In *Signal Processing and Communications Applications Conference (SIU), 2016 24th*.

Cem Üyük, Danica Rovó, Shaghayeghkolli Shaghayeghkolli, Rabia Varol, Georg Groh, and Daryna Dementieva. 2024. Crafting tomorrow's headlines: Neural news generation and detection in English, Turkish, Hungarian, and Persian. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 271–307, Miami, Florida, USA. Association for Computational Linguistics.

Oleg Vasilyev, Fumika Isono, and John Bohannon. 2024. Linear cross-lingual mapping of sentence embeddings. *Preprint*, arXiv:2305.14256.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. 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. *ArXiv*, abs/2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *ArXiv*, abs/2402.05672.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.

Seçil Özturk, Bülent Sankur, Tunga Gungör, Mustafa Berkay Yilmaz, Bilge Köroğlu, Onur Ağin, Mustafa İşbilen, Çağdaş Ulaş, and Mehmet Ahat. 2014. Turkish labeled text corpus. In *2014 22nd Signal Processing and Communications Applications Conference (SIU)*, pages 1395–1398.

Mert İncidelen and Murat Aydoğan. 2024. Developing question-answering models in low-resource languages: A case study on turkish medical texts using transformer-based approaches. In *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, pages 1–4.

A TR-MTEB Benchmark Datasets

All datasets included in the TR-MTEB benchmark were collected and used strictly for research purposes. Each dataset was accessed in accordance with its original license terms and usage conditions as specified on their respective repositories. We ensured that our use was consistent with the intended scope of these datasets, particularly when derivative resources (e.g., translated versions) were created. Furthermore, any models or corpora derived from these datasets are released under similar terms for non-commercial research use only, ensuring compliance with original data access conditions and ethical guidelines.

A.1 Classification

We compiled eight high-quality Turkish datasets to evaluate classification performance in the TR-MTEB benchmark. These datasets span a diverse set of domains and label categories:

- **THY Sentiment Classification**: A sentiment analysis dataset consisting of tweets collected from Turkish Airlines' official social media accounts. (Sezer, 2022)
- **TS Timeline News Category**: A news classification dataset containing articles from various Turkish online newspapers. (Sak et al., 2008)
- **75 News**: A small dataset with 75 Turkish news articles, equally distributed across five categories: economics, magazine, health, politics, and sports. (Amasyali and Yildirim, 2004)
- **Turkish Irony Classification**: An extended dataset for irony detection in Turkish social media posts. (Karagoz et al., 2021)
- **Turkish Movie Sentiment Classification**: A sentiment classification dataset with Turkish movie reviews. (Demirtas and Pechenizkiy, 2013)
- **Turkish News Category Classification**: A dataset with 1150 news articles annotated across five predefined categories. (Amasyalı and Yıldırım, 2004)
- **Turkish Offensive Language Classification**: Turkish portion of a multilingual offensive language dataset collected from Twitter. (Çöltekin, 2020)
- **Turkish Product Sentiment Classification**: Contains product reviews annotated with sentiment labels. (Demirtas and Pechenizkiy, 2013)

Table 3 summarizes these datasets with their sample sizes, language origin, and Hugging Face repository links:

Dataset Name	Train	Validation	Test	Language Origin	Hugging Face Link
THY Sentiment Classification	-	-	23.3k	Native Turkish	trmteb/thy_sa
TS Timeline News Category Classification	2k	-	2k	Native Turkish	trmteb/ts_timeline_news_category
75 News Classification	-	-	75	Native Turkish	trmteb/75haber
Turkish Irony Classification	-	-	600	Native Turkish	trmteb/irony-tr
Turkish Movie Sentiment Classification	7.97k	-	2.64k	Native Turkish	asparius/Turkish-Movie-Review
Turkish News Category Classification	750	150	250	Native Turkish	trmteb/news-cat
Turkish Offensive Language Classification	28k	3.28k	3.52k	Native Turkish	trmteb/offenseval
Turkish Product Sentiment Classification	4.8k	-	800	Native Turkish	asparius/Turkish-Product-Review

Table 3: Overview of datasets used in the classification task. The table includes sample sizes, language origin, and Hugging Face links.

A.2 Clustering

To evaluate the clustering capabilities of Turkish sentence embedding models, we selected two datasets that reflect distinct textual domains:

- **Turkish Abstract Corpus Clustering**: This dataset contains 6,234 academic paper abstracts categorized by scientific discipline. It was originally created to explore discipline-specific lexical patterns and language usage in academic writing. (Amasyalı and Diri, 2006)

- **Turkish Column Writing Clustering**: Comprises 630 Turkish opinion articles from 18 authors, with each author contributing 35 pieces. The dataset is designed to test the model's ability to cluster writings by authorship. (Özturk et al., 2014)

The statistics and metadata for these datasets are summarized in Table 4.

Dataset Name	Train	Validation	Test	Language Origin	Hugging Face Link
Turkish Abstract Corpus Clustering	-	-	6234	Native Turkish	trmteb/ts_abstract_corpus_p2p
Turkish Column Writing Clustering	-	-	630	Native Turkish	trmteb/630koseyazisi_p2p

Table 4: Overview of datasets used in the clustering task. The table includes sample sizes, language origin, and Hugging Face repository links.

A.3 Pair Classification

For evaluating sentence embedding models on pair classification tasks (e.g., natural language inference and paraphrase detection), we included three well-established datasets that were either translated into Turkish or contain a Turkish subset:

- **MNLI TR**: A machine-translated version of the Multi-Genre Natural Language Inference (MNLI) corpus. This dataset provides a large-scale resource for testing entailment relationships in Turkish. (Budur et al., 2020)

- **SNLI TR**: A Turkish translation of the Stanford Natural Language Inference (SNLI) dataset, also generated via neural machine translation (NMT). (Budur et al., 2020)

- **XNLI (Turkish Split)**: The Turkish subset of the Cross-lingual Natural Language Inference (XNLI) corpus, originally built to evaluate multilingual understanding and transfer. (Conneau et al., 2018)

Table 5 summarizes the key statistics and sources for these datasets.

Dataset Name	Train	Validation	Test	Language Origin	Hugging Face Link
MNLI TR	393k	10k	10k	Pre-Translated	trmteb/multinli_tr
SNLI TR	550k	10k	10k	Pre-Translated	trmteb/snli_tr
XNLI	393k	2.49k	5.01k	Pre-Translated	mteb/xnli

Table 5: Overview of datasets used in the pair classification task. The table includes sample sizes, language origin, and Hugging Face repository links.

A.4 Semantic Textual Similarity

To evaluate models on their ability to assess the semantic similarity between sentence pairs, we included the Turkish version of the STS Benchmark:

- **STSb TR**: This dataset is a machine-translated version of the original English STS benchmark. It was translated into Turkish using the Google Cloud Translation API without any post-editing. It provides graded similarity scores for sentence pairs and serves as a widely used benchmark in semantic similarity tasks. (Beken Fikri et al., 2021)

Table 6 provides key information about the dataset.

Dataset Name	Train	Validation	Test	Language Origin	Hugging Face Link
STSb TR	5.75k	1.5k	1.38k	Pre-Translated	trmteb/stsb-tr

Table 6: Overview of the dataset used in the semantic textual similarity task.

A.5 Bitext Mining

To evaluate cross-lingual alignment capabilities of embedding models, we included the English-Turkish portion of the WMT16 dataset:

- **WMT16 EN-TR**: This dataset comprises parallel sentence pairs in English and Turkish and is a subset of the well-known WMT16 machine translation benchmark. It serves as a reliable resource for evaluating sentence embeddings on bitext mining tasks by identifying translation equivalents between the two languages. (Bojar et al., 2016)

Table 7 summarizes the key metadata for the dataset.

Dataset Name	Train	Validation	Test	Language Origin	Hugging Face Link
WMT16 EN-TR	206k	1k	3k	Native Turkish	trmteb/wmt16_en_tr

Table 7: Overview of the dataset used in the bitext mining task.

A.6 Retrieval

To support evaluation of sentence embedding models on retrieval tasks, TR-MTEB includes a diverse collection of datasets across both general and domain-specific settings. These datasets span question answering, fact verification, and community-based QA. Below is a brief description of the included datasets:

- **SQuAD TR:** A machine-translated version of the original SQuAD v2.0 dataset using Amazon Translate, adapted for information retrieval tasks. (Budur et al., 2024).
- **TQuAD:** A native Turkish dataset focused on Turkish and Islamic Science History, developed for a national AI competition. (Soygazi et al., 2021).
- **XQuAD:** A cross-lingual QA benchmark adapted for retrieval, using the Turkish subset. (Artetxe et al., 2019).
- **MSMARCO TR:** A Turkish machine translation of the MSMARCO Passage Ranking dataset. (Kesgin et al., 2023).
- **NFCorpus TR:** A medical information retrieval corpus translated into Turkish as part of this study. (Boteva et al., 2016).
- **ArguAna TR:** A Turkish version of the ArguAna dataset created in this work. (Boteva et al., 2016).
- **CQA Dupstack Gaming TR:** A Turkish-translated subset of the CQADupStack benchmark focused on gaming-related community QA. (Hoogeveen et al., 2015).
- **FiQA 2018 TR:** A financial QA benchmark translated into Turkish in this study. (Thakur et al., 2021).
- **Quora TR:** A Turkish version of Quora duplicate question retrieval, translated during this work. (DataCanary, 2017).
- **SCIDOCS TR:** A document-level scientific benchmark translated into Turkish. (Cohan et al., 2020).
- **SciFact TR:** A Turkish-translated scientific fact verification dataset. (Wadden et al., 2022).

Table 8 summarizes the dataset properties.

Dataset Name	Corpus	Queries	Train	Val.	Test	Language Origin	Hugging Face Link
SQuAD TR	1.2k	8.29k	-	-	9.29k	Pre-Translated	trmteb/squad-tr
TQuAD	275	892	-	-	892	Native Turkish	trmteb/tquad
XQuAD	1.17k	1.19k	-	-	1.19k	Pre-Translated	google/xquad
MSMARCO TR	718k	501k	426k	53.3k	53.3k	Pre-Translated	trmteb/msmarco-tr
NFCorpus TR	3.63k	3.24k	111k	11.4k	12.3k	Translated in this study	trmteb/nfcorpus-tr
ArguAna TR	8.67k	1.41k	-	-	1.41k	Translated in this study	trmteb/arguana-tr
CQA Dupstack Gaming TR	45.3k	1.6k	-	-	2.26k	Translated in this study	trmteb/cqadupstack-gaming-tr
FiQA 2018 TR	57.6k	6.65k	14.2k	1.24k	1.71k	Translated in this study	trmteb/fiqa-tr
Quora TR	523k	15k	-	7.63k	15.7k	Translated in this study	trmteb/quora-tr
SCIDOCS TR	25.7k	1k	-	-	29.9k	Translated in this study	trmteb/scidocs-tr
SciFact TR	5.18k	1.11k	919	-	339	Translated in this study	trmteb/scifact-tr

Table 8: Overview of datasets used in the retrieval task.

B Translation of the Retrieval Dataset

B.1 Translation Prompt

```
Translate English texts from the BEIR benchmark dataset into Turkish. The purpose of
    this task is to create a Turkish benchmark using accurate translations that
    retain the context and meaning of the original texts.

# Steps
1. **Read the English Text**: Carefully read the provided English text from the BEIR
    dataset to fully understand its content and context.
2. **Translate to Turkish**: Convert the text into accurate and contextually
    appropriate Turkish, ensuring the meaning is preserved.
3. **Review and Adjust**: Review the translation for any errors or awkward phrasing,
    making adjustments to improve readability and accuracy.
4. **Finalize the Translation**: Confirm that the final translation accurately
    reflects the original English text in meaning and tone.

# Output Format
The output should be a well-structured Turkish translation, formatted in plain text.
    Each paragraph of the English text should have a corresponding Turkish
    translation, with clear delineation between sections if applicable.

# Notes
- Pay close attention to technical terms and ensure they are translated accurately.
- Maintain consistency in translation for recurring terms and phrases throughout the
    dataset.
- Consider the cultural context or expressions that may not have a direct
    translation into Turkish, and adapt them appropriately.

DO NOT ANSWER THE QUESTION IF QUESTION EXIST IN THE USER TEXT THAT WILL BE
    TRANSLATED. ONLY TRANSLATE THE TEXT. OUTPUT SHOULD BE ONLY THE TRANSLATED TEXT.

# English Text:

# Translated Text:
```

B.2 LLM-as-a-Judge Evaluation Prompt

```
You are an expert bilingual evaluator proficient in English and Turkish. Your task
    is to assess the quality of a machine-generated Turkish translation of an
    English text. The translation should be accurate, fluent, and contextually
    appropriate, preserving the original meaning, tone, and structure while ensuring
    natural readability in Turkish.

Evaluation Criteria:
Assess the translation holistically, considering:

Accuracy: Does the translation correctly convey the original meaning without
    omissions, distortions, or mistranslations?
Fluency: Is the Turkish text natural, grammatically correct, and free of awkward
    phrasing?
Terminology & Consistency: Are technical terms, domain-specific language, and
    recurring phrases translated correctly and consistently?
```

```
Cultural & Contextual Appropriateness: Does the translation adapt idioms,
    expressions, and culturally specific elements appropriately while maintaining
    the original intent?

Scoring Methodology (PASS/FAIL):
PASS     The translation is valid based on evaluation criteria above
FAIL     The translation is invalid based on evaluation criteria above and mostly
    fail.

A translation is considered PASS if it successfully conveys the intended meaning
    while maintaining fluency and accuracy. Otherwise, it is marked as FAIL.

Evaluation Task:
Analyze the following English text and its corresponding Turkish translation, then
    determine whether the translation passes or fails.

Input Format
English Text:
{original_text}

Translated Text:
{translated_text}

Output Format:
The output MUST be one of the PASS or FAIL in JSON format.

"PASS" means the translation meets quality standards.
"FAIL" means the translation does not meet quality standards.

Example Output:
{
    "decision": PASS/FAIL
}
```

C Evaluated Sentence Embedding Models

This section provides a detailed overview of all sentence embedding models evaluated in the TR-MTEB benchmark. The models span a range of architectures, languages, training regimes, and embedding strategies. We include both open-source multilingual and monolingual models, static word embeddings, and proprietary API-based models to ensure a comprehensive evaluation of sentence representation capabilities in Turkish.

C.1 Model Overview

The models can be categorized into the following groups:

- **Multilingual Transformers:** Large-scale pretrained models trained on multilingual corpora (e.g., E5, GTE, LaBSE).
- **SBERT-Based Transformers:** Sentence Transformers trained using contrastive objectives, covering both multilingual and English-only variants.
- **Monolingual Turkish Models:** BERT-based models pretrained on Turkish corpora, used both as baselines and for training our Turkish-specific embedding models.
- **Static Embeddings:** FastText and GloVe word embeddings for Turkish, aggregated into sentence embeddings by averaging.
- **API-Based Proprietary Models:** Embedding models accessible via API, used for benchmarking production-grade systems.

C.2 Model Specifications

Model specifications are summarized in Table 9.

Model Name	Size	Dim	Language	Reference	Accessible URL
multilingual-e5-large	560M	1024	Multilingual	(Wang et al., 2022)	intfloat/multilingual-e5-large
multilingual-e5-large-instruct	560M	1024	Multilingual	(Wang et al., 2024)	intfloat/multilingual-e5-large-instruct
gte-multilingual-base	305M	1024	Multilingual	(Li et al., 2023)	Alibaba-NLP/gte-multilingual-base
multilingual-e5-base	278M	768	Multilingual	(Wang et al., 2022)	intfloat/multilingual-e5-base
multilingual-e5-small	118M	384	Multilingual	(Wang et al., 2022)	intfloat/multilingual-e5-small
LaBSE	471M	768	Multilingual	(Feng et al., 2022)	sentence-transformers/LaBSE
paraphrase-multilingual-mpnet-base-v2	278M	768	Multilingual	(Reimers and Gurevych, 2019)	sentence-transformers/paraphrase-multilingual-mpnet-base-v2
paraphrase-multilingual-MiniLM-L12-v2	118M	768	Multilingual	(Reimers and Gurevych, 2019)	sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
all-MiniLM-L12-v2	33M	384	English	(Reimers and Gurevych, 2019)	sentence-transformers/all-MiniLM-L12-v2
all-MiniLM-L6-v2	22M	384	English	(Reimers and Gurevych, 2019)	sentence-transformers/all-MiniLM-L6-v2
all-mpnet-base-v2	109M	768	English	(Reimers and Gurevych, 2019)	sentence-transformers/all-mpnet-base-v2
bert-base-turkish-uncased	110M	768	Turkish	(Schweter, 2020)	dbmdz/bert-base-turkish-uncased
fasttext-turkish	-	300	Turkish	(Bojanowski et al., 2016)	fasttext.cc/docs/en/crawl-vectors.html
glove-turkish	-	-	Turkish	(Sarıtaş et al., 2024)	Turkish-Word-Embeddings/Word-Embeddings-Repository-for-Turkish
text-embedding-3-small	N/A	768	English (API)	(OpenAI, 2024)	models/text-embedding-3-small
turkish-embedding-model (ours)	110M	768	Turkish	This work	trmteb/turkish-embedding-model
turkish-embedding-model-fine-tuned (ours)	110M	768	Turkish	This work	trmteb/turkish-embedding-model-fine-tuned

Table 9: Overview of sentence embedding models used in TR-MTEB experiments.

D Turkish Sentence Embedding Training Data

To train robust and generalizable Turkish sentence embedding models, we compiled a large-scale weakly supervised dataset constructed from over 50 publicly available Turkish-language resources on the Hugging Face Hub. These datasets span a broad range of domains and formats, including question answering, summarization, instruction following, legal and medical dialogues, Wikipedia-based corpora, and user-generated content.

Table 10 provides a comprehensive overview of the data sources used to construct this corpus. Each entry includes metadata such as dataset type, language origin, original and filtered sentence pair counts, and dataset availability through Hugging Face. When applicable, academic references are provided.

To ensure domain and linguistic diversity, we prioritized datasets that were either native to Turkish or machine-translated with high quality. Some resources were synthetically generated using large language models or curated through human annotation pipelines. The dataset collection includes both formal and informal registers, and spans domains such as healthcare, finance, science, education, law, and conversational AI.

All datasets listed in this section were obtained from publicly available repositories (primarily Hugging Face) and used exclusively for research purposes. Each dataset was accessed in accordance with its stated license and terms of use. Where applicable, we respected the original access conditions, particularly for datasets derived from machine translation or synthetic generation. All derived artifacts, including the final sentence pair corpus and trained models, are intended solely for academic research and are distributed under non-commercial use terms, consistent with the sources from which they were built.

Table 10: Comprehensive list of Turkish language datasets used to construct the weakly supervised sentence pairs corpus for embedding model training. The table includes dataset types, origin, filtered pair counts, references (if available), and Hugging Face links.

Dataset Name	Dataset Type	Language Origin	Pair Count	Filtered Pair Count	Reference (if any)	Hugging Face Link
WikiRAG-TR	Q&A	Synthetically Generated	18k	9.73k	-	Metin/WikiRAG-TR
wikisource	Title&Text	Native Turkish	7.22k	4.07k	(Foundation)	wikimedia/wikisource

Table 10: Comprehensive list of Turkish language datasets used to construct the weakly supervised sentence pairs corpus for embedding model training. The table includes dataset types, origin, filtered pair counts, references (if available), and Hugging Face links (cont.).

Dataset Name	Dataset Type	Language Origin	Pair Count	Filtered Pair Count	Reference (if any)	Hugging Face Link
mqa	Q&A	Native Turkish	6.66M	3.79M	(De Bruyn et al., 2021)	clips/mqa
mfaq	Q&A	Native Turkish	102k	61.1k	(De Bruyn et al., 2021)	clips/mfaq
tapaco	Paraphrase	Native Turkish	183k	168k	(Scherrer, 2020)	community-datasets/tapaco
Turkish Medical Reasoning	Q&A with Content	Translated	7.21k	6.96k	-	iuperceptron/ turkish_medical_reasoning
wiki_lingua	Title&Content	Native Turkish	4.5k	2.73k	(Ladhak et al., 2020)	esdurmus/wiki_lingua
oasst1 Pairwise RLHF Reward	Q&A	Instruction	9	5	-	tasksource/ oasst1_pairwise_rlhf_reward
Wikipedia TR Summarization	Summarization	Native Turkish	125k	125k	(Gultekin, 2023)	musabg/wikipedia-tr-summarization
Turkish Instructions	Instruction	Translated	51.9k	42k	-	SoAp9035/turkish_instructions
Wikinews Multilingual	Title&Content	Native Turkish	363	252	(Vasilyev et al., 2024)	Fumika/Wikinews-multilingual
OpenOrca TR	Instruction	Translated	798k	649k	(Lian et al., 2023)	ucekmez/OpenOrca-tr
ARC TR	Q&A	Translated	31.1k	4.41k	(Clark et al., 2018)	malhajar/arc-tr
OO GPT4 Filtered TR	Q&A	Translated	101k	85.8k	-	umarigan/oo-gpt4-filtered-tr
babel briefings	Q&A with Context	Native Turkish	341k	260,1k	(Leeb and Schölkopf, 2024)	felixludos/babel-briefings
flickr8k TR	Paraphrase	Synthetically Generated	8k	6.48k	(Unal et al.)	atasoglu/flickr8k-turkish
SE dataset	Instruction	Native Turkish	877	765	-	salihturkoglu/se_data_set
Turkish Youtube Comments	Comment&Reply	Native Turkish	8.91k	3.49k	-	yusiqo/Turkish-Youtube-Comments
Turkish Medical QA	Q&A	Native Turkish	188k	75.1k	(Bayram, 2024c)	alibayram/doktorsitesi
Onedio Haberler	Title&Content	Native Turkish	66.7k	59.5k	(Bayram, 2024a)	alibayram/onedio_haberler
CC News	Title&Content	Native Turkish	3.72M	3.25M	-	stanford-oval/ccnews
Neural News Benchmark	Title&Content	Native Turkish	3.21k	1.92k	(Üyük et al., 2024)	tum-nlp/neural-news-benchmark
gsm8k TR	Q&A	Translated	8.76k	8.74k	(Bezir, 2024)	bezir/gsm8k-tr
InstrucTurca	Instruction	Synthetically Generated	363	252	(Altinok, 2024)	turkish-nlp-suite/InstrucTurca
Turkish Law Chatbot	Q&A	Native Turkish	14.9k	12.5k	-	Renicames/turkish-law-chatbot
Wikipedia 2024-06 bge-m3	Title&Content	Native Turkish	2.05M	1.25M	-	Upstash/wikipedia-2024-06-bge-m3
alpaca TR	Instruction	Translated	45.3k	27.7k	-	BrewInteractive/alpaca-tr
Turkish LLM Finetune Dataset	Q&A with Context	Native Turkish	193.6k	150.9k	-	barathanasln/turkish_llm _finetune_dataset_4_topics
TRT Data Warrriors Dataset	Q&A	Native Turkish	2k	1.97k	-	NusretOzates/ TRTDataWarrriorsDataset
Multilingual Reward Bench	Instruction	Native Turkish	2.87k	2.59k	(Gureja et al., 2024)	CohereLabsCommunity/ multilingual-reward-bench
TDK Sozluk Turkish	Paraphrase	Native Turkish	133k	72.8k	-	Ba2han/TDK_Sozluk-Turkish
turkish exam instructions	Instruction	Native Turkish	41.4k	39.9k	(Bayram, 2024b)	bezir/turkish_exam_instructions
Wikipedia Turkish QA Chat	Q&A	Native Turkish	1.6M	784k	-	Quardo/wikipedia-turkish-qa-chattemplate
News TR	Title&Content	Native Turkish	1.53M	1.36M	-	habanoz/news-tr-1.8M
MedTurkQuAD	Q&A with Context	Native Turkish	24.6k	7.82k	(İncidelen and Aydoğan, 2024)	incidelen/MedTurkQuAD
Legal NLI TR	Paraphrase	Native Turkish	203k	192k	-	Turkish-NLI/legal_nli_TR_V1
turoqa small	Instruction	Native Turkish	3k	3k	-	SoAp9035/turoqa-small
Turkish Wikipedia QA	Q&A	Native Turkish	4.01M	2.88M	-	hcsolakoglu/turkish-wikipedia-qa-4-million
ThinkingData Turkish	Instruction	Translated	209k	202k	-	erythropygia/ThinkingData-200K-Turkish
Multilingual NLI	Paraphrase	Translated	33.82k	26.62k	(Laurer et al., 2022)	MoritzLaurer/multilingual-NLI-26lang-2mil7
Seahorse Summarization	Text&Summary	Synthetically Generated	696.7k	310k	(Clark et al., 2023)	tasksource/ seahorse_summarization_evaluation
lr-sum	Text&Summary with Title	Synthetically Generated	107.4k	87.6k	(Palen-Michel and Lignos, 2023)	bltlab/lr-sum
LlamaTurk Instruction Set	Instruction	Translated	52k	41.6k	-	metunlp/LlamaTurk-Instruction-Set
xMINDlarge	Text&Summary	Native Turkish	279k	202k	(Iana et al., 2024)	aiana94/xMINDlarge
Patient-Doctor QA	Q&A	Translated	234.9k	189.1k	(Bulut, 2024)	kayrab/patient-doctor-qa-tr-19583
zynp_ai teknofest	Q&A	Native Turkish	35M	14M	(sekerlipencere, 2024)	sekerlipencere/zynpdata-zynp_ai-teknofest
Turkish Headline Dataset	Text&Summary	Native Turkish	14.6k	12.2k	-	nuilbg/turkish_headline_dataset_new
MLSUM	Summarization	Native Turkish	822k	721k	(Scialom et al., 2020)	reciTAL/mlsum

Table 10: Comprehensive list of Turkish language datasets used to construct the weakly supervised sentence pairs corpus for embedding model training. The table includes dataset types, origin, filtered pair counts, references (if available), and Hugging Face links (cont.).

Dataset Name	Dataset Type	Language Origin	Pair Count	Filtered Pair Count	Reference (if any)	Hugging Face Link
TRSUM	Summarization	Native Turkish	921k	827k	(Baykara and Güngör, 2022)	-
Turkish Labeled Text Corpus	Summary	Native Turkish	7.9k	6.61k	(Özturk et al., 2014)	-

E Training Details

We summarize the hyperparameters and compute settings used for both the contrastive pretraining and supervised fine-tuning stages in Table 11.

Setting	Contrastive Pretraining	Supervised Fine-Tuning
Model Initialization	`dbmdz/bert-base-turkish-uncased`	Contrastively pretrained checkpoint
Batch Size (per device)	32,768 (effective, via caching)	64
Learning Rate	3e-5	2e-5
Optimizer	AdamW (default in `SentenceTransformers`)	AdamW
Warmup Ratio	0.05	0.1
Loss Function	Cached-MNRL	Task-specific (classification/retrieval) loss
Epochs	1	1
Precision	bfloat16 (bf16)	bfloat16 (bf16)
Evaluation Strategy	Every 100 steps	Every 1000 steps
Save Strategy	Every 100 steps	Every 1000 steps
Compute Resources	1 $\times$ NVIDIA A100 (40GB)	1 $\times$ NVIDIA A100 (40GB)
Approx. Training Duration	$\sim$80 hours	$\sim$2 hours

Table 11: Summary of training hyperparameters and compute resources for contrastive pretraining and supervised fine-tuning.