TR-MTEB: A COMPREHENSIVE BENCHMARK AND EMBEDDING MODEL SUITE FOR TURKISH SENTENCE REPRESENTATIONS

by

Mehmet Selman Baysan

B.S., Computer Engineering, Marmara University, 2023

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2025

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my thesis advisor, Prof. Tunga Güngör, for his invaluable guidance, encouragement, and support throughout this journey. His mentorship has been instrumental in shaping the direction and quality of this work.

I would also like to extend my sincere thanks to Assoc. Prof. Murat Can Ganiz, both a member of my thesis committee and my undergraduate advisor. He was the one who first introduced me to this field and played a pivotal role in inspiring my early academic pursuits. I am truly grateful for the lasting impact he has had on my academic and professional path.

My heartfelt appreciation also goes to Assist. Prof. Suzan Üsküdarlı, a member of my thesis committee, for her valuable feedback and contributions to this thesis.

On a more personal note, I would like to thank all my friends, far too many to name here, for their unwavering emotional support and patience during this process. In particular, I am grateful to Mehmet Ulutaş, Hüseyin Selim Çoban, Fatih Satı, Salih Gökçe, and Ayfer Torunoğlu for always being there when I needed them most.

Finally, I am profoundly thankful to my mother, father, and brother for their unconditional love, support, and belief in me. This achievement would not have been possible without them.

ABSTRACT

TR-MTEB: A COMPREHENSIVE BENCHMARK AND EMBEDDING MODEL SUITE FOR TURKISH SENTENCE REPRESENTATIONS

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 integrates 26 high-quality datasets, including both native Turkish corpora and carefully curated machine-translated resources.

To support embedding model development, we construct a 34.2 million pair Turkish sentence corpus derived from over 50 public datasets, applying multi-stage filtering to ensure semantic alignment. Using this corpus, we train two monolingual Turkish sentence embedding models from a `BERTurk` backbone: one with contrastive pretraining using Cached-MNRL and one further fine-tuned on TR-MTEB tasks.

Despite being trained with significantly less data and computation resources than their multilingual counterparts, our models outperform the Turkish BERT baseline across all tasks and achieve competitive performance in classification and similarity, while highlighting limitations in cross-lingual alignment.

All datasets, models, and code are publicly released to promote reproducibility and further research in Turkish natural language processing and low-resource sentence representation.

ÖZET

TR-MTEB: TÜRKÇE CÜMLE TEMSİLLERİ İÇİN KAPSAMLI BİR DEĞERLENDİRME SETİ VE VEKTÖREL TEMSİL MODEL PAKETİ

Bu çalışmada, Türkçe cümle temsil modellerini değerlendirmek üzere tasarlanmış, görev çeşitliliği yüksek ilk büyük ölçekli ölçüt seti olan TR-MTEB geliştirilmiştir. Sınıflandırma, kümeleme, ikili sınıflandırma, bilgi getirme, iki dilli metin eşleştirme ve anlamsal metin benzerliği olmak üzere altı temel görevi kapsayan TR-MTEB, hem özgün Türkçe veri kümelerini hem de özenle seçilmiş makine çevirisi kaynaklarını içeren toplam 26 yüksek kaliteli veri kümesini bir araya getirmektedir.

Model geliştirmeyi desteklemek amacıyla, 50'nin üzerinde açık veri kümesinden türetilmiş ve anlamsal uyumu sağlamak için çok aşamalı filtreleme uygulanmış 34,2 milyon cümle çiftinden oluşan bir Türkçe cümle korpusu oluşturulmuştur.Bu derlem kullanılarak, `BERTurk` tabanlı iki tek dilli Türkçe cümle temsil modeli eğitilmiştir: biri Cached-MNRL yöntemiyle karşıtlık temelli ön eğitim ile hazırlanmış, diğeri ise TR-MTEB görevleri üzerinde ince ayar yapılmış bir modeldir.

Bu modeller, çok dilli karşılaştırmalı modellerden çok daha az veri ve hesaplama kaynağıyla eğitilmelerine rağmen, tüm görevlerde Türkçe BERT tabanlı modeli geride bırakmış; özellikle sınıflandırma ve benzerlik görevlerinde benzer performans sergilemiş, ancak çapraz dilli hizalama konusunda mevcut sınırlamaları da ortaya koymuştur.

Tüm veri kümeleri, modeller ve kodlar açık olarak paylaşılmış olup, Türkçe doğal dil işleme ve düşük kaynaklı dil temsilleri alanında yeniden üretilebilirliği ve ileri araştırmaları teşvik etmeyi amaçlamaktadır.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . iii
ABSTRACT . . . iv
ÖZET . . . v
LIST OF FIGURES . . . ix
LIST OF TABLES . . . x
LIST OF SYMBOLS . . . xiii
LIST OF ACRONYMS/ABBREVIATIONS . . . xiv
1. INTRODUCTION . . . 1
2. RELATED WORK . . . 4
2.1. The Emergence of Massive Text Embedding Benchmarks . . . 4
2.2. Language-Specific Adaptations of MTEB . . . 5
2.3. Cross-Lingual Retrieval Extensions of the BEIR Benchmark . . . 8
2.4. General Text Embedding Models . . . 9
2.4.1. Large-Scale Sentence Pair Corpora . . . 10
2.4.2. Training Universal Embedding Models . . . 11
3. METHODOLOGY . . . 13
3.1. Turkish Massive Text Embedding Benchmark (TR-MTEB) . . . 14
3.1.1. Tasks Overview . . . 14
3.1.1.1. Classification . . . 15
3.1.1.2. Clustering . . . 15
3.1.1.3. Pair Classification . . . 15
3.1.1.4. Bitext Mining . . . 15
3.1.1.5. Semantic Textual Similarity (STS) . . . 16
3.1.1.6. Retrieval . . . 16
3.1.2. Collecting and Structuring TR-MTEB Datasets . . . 17
3.1.2.1. Classification . . . 17
3.1.2.2. Clustering . . . 19
3.1.2.3. Pair Classification . . . 21

3.1.2.4. Semantic Textual Similarity (STS) 22
3.1.2.5. Bitext Mining 23
3.1.2.6. Retrieval 24
3.1.3. Translation of Retrieval Datasets from BEIR Benchmark 26
3.1.3.1. LLM-based Translation Strategy 27
3.1.3.2. Translation Evaluation: LLM-as-a-Judge 28
3.1.3.3. Post-Editing Pipeline 31
3.1.3.4. Downstream Use 31
3.2. Large-Scale Turkish Sentence Pair Corpus 32
3.2.1. Training Turkish Sentence Embedding Models 43
3.2.1.1. Contrastive Pretraining with Cached-MNRL 43
3.2.1.2. Supervised Fine-Tuning 44
3.2.1.3. Training Configuration 45
3.2.1.4. Evaluation Pipeline 45
3.2.1.5. Our Models 46
4. EXPERIMENTS AND RESULTS 47
4.1. Model Selection Strategy and Model Specifications 47
4.2. Performance Analysis and Evaluation Insights 52
4.2.1. Global Trends in Mean(Task) Performance 56
4.2.2. Performance Across Task Types 57
4.2.2.1. Classification Tasks 57
4.2.2.2. Clustering Tasks 57
4.2.2.3. Bitext Mining Task 57
4.2.2.4. Pair Classification Tasks 58
4.2.2.5. Semantic Textual Similarity (STS) Task 58
4.2.2.6. Retrieval Tasks 58
4.2.3. Impact of Contrastive Pretraining 58
4.2.4. Effectiveness of Supervised Fine-Tuning 59
4.2.5. Architectural and Capacity-Based Comparisons 59
4.2.6. Cross-Lingual Alignment Limitations 59
4.2.7. Analysis of Score Variability 60

4.2.8. Summary of Findings 60
4.3. Limitations 61
5. CONCLUSION 63
REFERENCES 66
APPENDIX A: MTEB-TR Task Results 75
A.1. Classification Results 75
A.2. Clustering Results 75
A.3. Bitext Mining Results 75
A.4. Pair Classification Results 75
A.5. Semantic Textual Similarity Results 76
A.6. Retrieval Results 76

LIST OF FIGURES

Figure 3.1. An overview of tasks and datasets included in TR-MTEB. Datasets in white background are original Turkish datasets. Datasets highlighted in green indicate those that were translated to Turkish prior to this study, while datasets marked in blue represent English datasets that were translated into Turkish as part of this work. 13

Figure 3.2. Prompt for Translating BEIR Dataset into Turkish 28

Figure 3.3. Confusion matrix comparing human annotations and LLM predictions for translation quality evaluation. 29

Figure 3.4. Prompt template for LLM-as-a-Judge translation evaluation . . . 30

Figure 3.5. Distribution of filtered sentence pairs by dataset type. For visualization purposes, specific dataset types were grouped into broader categories such as *Question & Answer*, *Title & Content*, *Instruction*, etc. This aggregation simplifies presentation and enhances interpretability while reflecting the core structural formats used for weak supervision. 41

Figure 4.1. Overall Mean(Task) performance of sentence embedding models on TR-MTEB. 53

Figure 4.2. Distribution of model performance across six TR-MTEB task categories. Each box shows the score spread across models per task. . 61

LIST OF TABLES

Table 3.1. Overview of classification datasets in TR-MTEB. All datasets are in Turkish and span diverse domains including social media, e-commerce, and news media. 19

Table 3.2. Overview of datasets used in the clustering task. Both datasets are used in a zero-shot setting with mini-batch K-Means and evaluated via V-Measure. 20

Table 3.3. Overview of datasets used in the pair classification task. Labels were binarized to distinguish between `entailment` and `non-entailment`. 22

Table 3.4. Overview of the dataset used in the semantic textual similarity task. Cosine similarity between sentence embeddings is evaluated using Spearman and Pearson correlations against the annotated similarity scores. 23

Table 3.5. Overview of the dataset used in the bitext mining task. The dataset contains professionally translated English–Turkish sentence pairs and is used to assess cross-lingual semantic alignment capabilities of embedding models. 24

Table 3.6. Overview of datasets used in the retrieval task. Corpus and query sizes are provided along with data split availability, translation status, and Hugging Face links. 27

Table 3.7: 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. 33

Table 3.8. 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 4.2. 42

Table 4.1: Overview of sentence embedding models used in TR-MTEB experiments, extended with context length. 48

Table 4.2: TR-MTEB benchmark results across embedding models. Bitext: Bitext Mining, Class.: Classification, Clust.: Clustering, Pair Cl.: Pair Classification, Retr.: Retrieval, STS: Semantic Textual Similarity. Results ordered by Mean (Task) scores. 53

Table A.1. Classification task performance across models in TR-MTEB. Scores reflect macro-average F1 across categories. 77

Table A.2. Performance of embedding models on Turkish clustering tasks. Evaluation based on V-measure for two domains: academic abstracts and opinion columns. 78

Table A.3. Performance of embedding models on Turkish–English bitext mining using the WMT16 EN-TR dataset. Evaluation based on F1-score for top-1 retrieval. 79

Table A.4. Performance of embedding models on pair classification tasks in Turkish using XNLI, SNLI TR, and MultiNLI-TR datasets. Evaluation uses macro-average F1 over entailment detection. 80

Table A.5. Semantic similarity performance of embedding models on the Turkish version of the STS-Benchmark dataset. Evaluation based on Spearman correlation. 81

Table A.6. Retrieval task results on Turkish or previously translated datasets such as MS MARCO TR, SQuAD TR, TQuAD, and XQuAD. Evaluation is based on nDCG@10. 82

Table A.7. Retrieval task results on Turkish versions of BEIR datasets translated in this work. Includes ArguAna TR, FiQA TR, CQADupstack Gaming TR, NFCorpus TR, and others. Evaluation metric is nDCG@10. 83

LIST OF SYMBOLS

a_i	Anchor embedding in the i-th sentence pair
$\hat{y}_i$	Predicted class probability for instance i
n	Batch size or total number of instances
p_i	Positive (paired) embedding in the i-th sentence pair
$\mathbf{u}_i$	First sentence embedding in a sentence pair (for STS)
$\mathbf{v}_i$	Second sentence embedding in a sentence pair (for STS)
y_i	Ground-truth label or similarity score for instance i
α	Temperature scaling parameter in CoSENT loss
$\mathcal{L}_{\text{CLS}}$	Cross-entropy loss for classification tasks
$\mathcal{L}_{\text{CoSENT}}$	CoSENT loss for semantic similarity tasks
$\mathcal{L}_{\text{MNRL}}$	Multiple Negatives Ranking Loss for contrastive training
τ	Temperature parameter in MNRL formulation

LIST OF ACRONYMS/ABBREVIATIONS

API	Application Programming Interface
BEIR	Benchmarking Information Retrieval
BERT	Bidirectional Encoder Representations from Transformers
BERTurk	BERT model trained on Turkish corpora
BGE	BAAI General Embeddings
CLS	Classification
C-MTEB	Chinese Massive Text Embedding Benchmark
CoSENT	Cosine Sentence Similarity Loss
E5	EmbEddings from bidirEctional Encoder rEpresentations
F1	F1-score (harmonic mean of precision and recall)
GPU	Graphics Processing Unit
GTE	General Text Embeddings
IR	Information Retrieval
LLM	Large Language Model
MNRL	Multiple Negatives Ranking Loss
MTEB	Massive Text Embedding Benchmark
MTP	Massive Text Pair
nDCG@10	normalized Discounted Cumulative Gain at rank 10
NLI	Natural Language Inference
NLP	Natural Language Processing
PL-MTEB	Polish Massive Text Embedding Benchmark
Ru-MTEB	Russian Massive Text Embedding Benchmark
SEB	Scandinavian Embedding Benchmark
SNLI	Stanford Natural Language Inference
STS	Semantic Textual Similarity
TR-MTEB	Turkish Massive Text Embedding Benchmark
XNLI	Cross-lingual Natural Language Inference

1. INTRODUCTION

Text embeddings lie at the heart of modern Natural Language Processing (NLP), enabling scalable and efficient representations for tasks such as document retrieval, semantic search, clustering, and classification. In recent years, universal sentence embedding models have become increasingly important due to their ability to generalize across domains and tasks with minimal supervision. While these models have shown impressive results in high-resource languages, especially English, their performance and development in morphologically rich and syntactically complex low-resource languages, such as Turkish, remain understudied and under-supported.

One of the major catalysts behind recent progress in embedding evaluation is the Massive Text Embedding Benchmark (MTEB) (1), which offers a multi-task evaluation suite covering classification, clustering, retrieval, and more. Inspired by MTEB, language-specific extensions like C-MTEB for Chinese (2) and PL-MTEB for Polish (3) have demonstrated that adapting benchmarks to linguistic, cultural, and structural characteristics of individual languages results in more accurate and meaningful evaluations. However, Turkish, despite being spoken by tens of millions and widely used online, lacks such a comprehensive resource. At the time of writing, Turkish is represented in the multilingual MTEB benchmark by only a few datasets and tasks, severely limiting both evaluation granularity and the training of Turkish-specific models.

To address this gap, we introduce the Turkish Massive Text Embedding Benchmark (TR-MTEB), the first standardized benchmark for evaluating Turkish sentence embeddings across six NLP task categories: *classification, clustering, pair classification, retrieval, bitext mining*, and *semantic textual similarity (STS)*. The benchmark comprises a total of 26 datasets, combining both native Turkish resources and high-quality machine-translated English benchmarks. Notably, the retrieval subset is extended by translating seven widely used datasets from the BEIR benchmark (4), with translation quality ensured through an LLM-as-a-Judge evaluation pipeline and manual

corrections.

Beyond evaluation, we tackle the data scarcity problem for training Turkish embedding models. Existing multilingual models such as E5 (5), GTE (6), and BGE (2) have benefited from hundreds of millions of multilingual sentence pairs. In contrast, Turkish lacks any comparable large-scale sentence pair resource. To fill this void, we curated a high-quality weakly supervised corpus of 34.2 million Turkish sentence pairs. These pairs were collected from over 50 publicly available Turkish datasets hosted on the Hugging Face Hub [1] , covering a wide range of domains such as education, healthcare, news, user-generated content, and Wikipedia. The raw corpus was processed using a multi-stage filtering pipeline based on similarity scores computed by multilingual embedding models, and further refined using a fine-tuned E5 encoder to ensure semantic relevance.

To demonstrate the effectiveness of this corpus, we trained two Turkish-specific sentence embedding models using the monolingual `bert-base-turkish-uncased` architecture (7). The first model was trained using contrastive learning with Cached Multiple Negatives Ranking Loss (MNRL) to simulate large batch sizes under resource constraints. The second model was fine-tuned using labeled TR-MTEB datasets across multiple downstream tasks. Our evaluation shows that both models significantly outperform the untuned Turkish BERT baseline and yield competitive results across most TR-MTEB tasks, especially in classification and semantic similarity. Although monolingual models underperform on cross-lingual tasks such as bitext mining, they provide strong baselines for Turkish NLP and highlight the viability of high-quality, language-specific training even under limited compute and data conditions.

Our contributions are: (1) TR-MTEB, the first comprehensive Turkish benchmark for evaluating sentence embeddings across six diverse NLP tasks; (2) a newly constructed large-scale corpus of 34.2M Turkish sentence pairs filtered and structured for embedding training; (3) two Turkish-specific embedding models trained using con-

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

trastive and supervised objectives; and (4) the public release of all datasets, models, and evaluation tools to foster future research in Turkish NLP and promote fair benchmarking practices for low-resource languages.

2. RELATED WORK

2.1. The Emergence of Massive Text Embedding Benchmarks

Text embeddings form the semantic backbone of many modern NLP systems, serving as universal representations of text that enable transfer learning across tasks such as classification, retrieval, clustering, and question answering. Traditionally, the evaluation of these representations was confined to a limited set of intrinsic tasks (e.g., word similarity) or sentence-level semantic similarity tasks (e.g., STS Benchmark). While useful, such evaluations offered only a narrow lens through which to judge the generalization and utility of embedding models, especially for real-world applications.

To address this limitation, the Massive Text Embedding Benchmark (MTEB) was proposed by Muennighoff et al. (1) as a comprehensive benchmark suite for evaluating general-purpose embedding models across a broad spectrum of NLP tasks. MTEB assesses model performance on eight core task types: classification, clustering, pair classification, reranking, retrieval, bitext mining, semantic textual similarity (STS), and summarization. It includes 58 datasets spanning 112 languages, encompassing both high-resource and low-resource languages.

MTEB distinguishes itself from earlier benchmarks like SentEval (8) or STS-B (9) in several key ways:

- Task and domain diversity: MTEB evaluates embeddings on a variety of tasks that require both symmetric and asymmetric semantic reasoning. For instance, symmetric tasks such as Semantic Textual Similarity require assessing the degree of equivalence between two sentences, whereas asymmetric tasks like retrieval focus on evaluating the relevance of a document given a query, which involves directional semantic matching.
- Zero-shot utility and modularity: Tasks in MTEB are evaluated in a zero-shot

or few-shot fashion, relying directly on the quality of the embeddings rather than task-specific finetuning. This design allows for clear insights into a model's transfer capabilities.

- Benchmark extensibility: Leveraging the Hugging Face Datasets and Metrics libraries, MTEB is designed to be extensible, new datasets or tasks can be easily integrated and standardized scripts ensure reproducibility.
- Public leaderboard and ecosystem support: MTEB includes a maintained leaderboard featuring over 30 embedding models such as SimCSE, SGPT, GTE, LaBSE, and MPNet, providing comparative insights into architecture, size, and training objectives.

Despite its multilingual ambitions, MTEB remains limited in its coverage of certain languages. For instance, Turkish, a morphologically rich and syntactically flexible language, is represented in the original MTEB with only three datasets. This underrepresentation not only restricts the development of robust Turkish-specific models but also impairs the fair evaluation of multilingual systems in low-resource contexts.

To bridge this gap, our work proposes TR-MTEB, a comprehensive Turkish extension of MTEB. TR-MTEB expands the original framework by incorporating 26 Turkish datasets across six task types, combining native resources and high-quality translations. Following the principles established by MTEB, task modularity, zero-shot evaluation, and format standardization, TR-MTEB offers a rigorous and extensible benchmark that is specifically tailored to the linguistic and resource realities of Turkish NLP.

2.2. Language-Specific Adaptations of MTEB

Following the success of the original MTEB framework (1), several language-specific adaptations have been proposed to address the unique syntactic, morphological, and resource characteristics of individual languages. These adaptations aim to provide more accurate evaluations for monolingual and multilingual models in underrepresented languages, often by incorporating native datasets or carefully translated resources, and

by adjusting task formulations to reflect linguistic realities.

C-MTEB (2) extends MTEB for Chinese by combining translated and native datasets to evaluate Chinese sentence embeddings across six task types. A notable contribution is the focus on instruction-tuned embedding models for Chinese, including prompts tailored for cultural nuances and character-level semantics. This adaptation shows how incorporating language-specific pretraining data and task instruction alignment improves evaluation fidelity in ideographic languages.

PL-MTEB (3) presents a Polish version of MTEB by introducing a diverse suite of native classification, retrieval, and STS datasets. The benchmark emphasizes high-quality preprocessing and scoring consistency, adapting STS and classification tasks to the nuances of Polish morphology. Notably, PL-MTEB includes task reformulations that better reflect real-world Polish NLP applications, such as question answering and public sector information retrieval.

MTEB-FR (10) adapts MTEB for French by translating core MTEB tasks and curating additional French-specific benchmarks in classification and retrieval. It further integrates sentence-level and document-level evaluations, allowing for a multi-granularity performance assessment. Importantly, MTEB-FR examines cross-lingual transfer effects by benchmarking French monolingual models against multilingual encoders, illustrating the trade-off between model scale and specialization.

Ru-MTEB (11) contributes a robust Russian embedding benchmark with over 20 datasets, including both general-domain and specialized corpora such as legal and scientific texts. The methodology focuses on preserving grammatical features inherent to Russian, such as aspectual verb forms and case declensions, during translation and evaluation. Ru-MTEB also explores sentence retrieval performance using culturally grounded information sources like Russian Wikipedia and news media.

Scandinavian Embedding Benchmark (SEB) (12) addresses the lack of reliable

evaluation frameworks for Danish, Swedish, and Norwegian embeddings. Unlike most prior adaptations, SEB explicitly avoids translated datasets wherever possible, prioritizing native corpora to ensure cultural and linguistic integrity. It covers 24 tasks across 10 subtasks and four core task categories: classification, clustering, retrieval, and bitext mining.

SEB also introduces several methodological and infrastructural enhancements that set it apart from prior adaptations. First, it provides a comprehensive model registry, facilitating reproducibility and comparative analysis across different embedding approaches. Moreover, the benchmark employs task-specific embedding strategies, such as using distinct prompts for classification and retrieval tasks, thereby aligning the evaluation setup more closely with real-world use cases. SEB further incorporates cross-lingual transfer evaluations, leveraging the high degree of mutual intelligibility among Scandinavian languages to examine model generalizability. Finally, it includes a visual analysis of embedding space similarity, enabling a qualitative inspection of task diversity and embedding behavior. These contributions collectively elevate SEB as a reference benchmark for low-resource languages, demonstrating how both linguistic fidelity and technical scalability can be achieved in embedding evaluations.

The TR-MTEB framework proposed in this work extends these prior efforts by introducing the first large-scale, Turkish-specific adaptation of the MTEB benchmark. Turkish presents unique linguistic challenges stemming from its agglutinative morphology and high information density at the sentence level. To address these, TR-MTEB incorporates a combination of native and carefully translated datasets spanning six core task types. The evaluation protocol adheres to MTEB standards while adapting scoring functions and sampling strategies to better reflect the syntactic and semantic characteristics of Turkish. Additionally, we train compact yet competitive Turkish-specific embedding models, providing strong baselines tailored to low-resource language contexts. Collectively, these language-specific extensions of MTEB—including our own—underscore the flexibility of the MTEB framework and emphasize the necessity of localized evaluation strategies. While grounded in a shared evaluation philosophy, each

adaptation introduces methodological innovations—ranging from dataset construction and translation validation to model training and task curation—demonstrating the value of language-aware benchmarking in advancing general-purpose embedding models.

2.3. Cross-Lingual Retrieval Extensions of the BEIR Benchmark

The BEIR benchmark (4) is a prominent resource for evaluating zero-shot information retrieval (IR) systems and serves as the foundation for evaluating Turkish retrieval capabilities in this study. Originally designed to benchmark dense retrievers on 18 English datasets across diverse domains—including question answering, fact checking, passage retrieval, and argument retrieval, BEIR provides well-established IR tasks that can be systematically adapted into new languages.

To address the benchmark's monolingual limitations, recent initiatives have translated BEIR into additional languages. One such initiative is BEIR-PL (13), which targets Polish, a morphologically rich and underrepresented language in IR. BEIR-PL involved machine translation of 13 BEIR datasets using Google Translate, followed by semantic and professional linguistic validation of a sample set to assess translation fidelity. The authors also contributed new Polish-language IR models, trained using dense retrievers and rerankers, and evaluated performance against strong lexical baselines like BM25. Their experiments showed significant gains from neural models in zero-shot settings, despite Polish's linguistic complexity and the resource constraints of lexical retrieval approaches.

Similarly, BEIR-NL (14) extends BEIR to Dutch using Gemini-based machine translation with prompt customization based on query/document type and domain context. BEIR-NL covers 14 datasets and introduces rigorous translation quality checks, including manual annotation of sampled translations for major and minor semantic issues. Additionally, the benchmark evaluates various multilingual dense retrievers and rerankers, emphasizing the challenges posed by translation-induced data shifts

and potential model-data contamination. The authors also conducted back-translation experiments, revealing measurable degradation in performance due to translation artifacts, which reinforces the importance of careful curation and evaluation in cross-lingual adaptations.

In our work, we adopt a similar methodology by translating seven BEIR datasets into Turkish, thereby creating a Turkish retrieval suite within the broader TR-MTEB. Beyond automated translation, we introduce a rigorous quality assurance pipeline involving LLM-as-a-Judge semantic scoring and targeted manual correction to ensure that the translated corpora preserve the original intent and retrieval relevance. This curated resource forms a key component of the retrieval task in TR-MTEB, supporting robust evaluation of Turkish language embedding models.

These multilingual extensions of BEIR not only demonstrate the feasibility of bootstrapping retrieval benchmarks in low-resource languages via high-quality translation and semantic validation, but also highlight the limitations of purely automated pipelines. As shown by both BEIR-PL and BEIR-NL, translation-induced semantic drift, lexical mismatches, and domain-specific challenges can impact IR performance, underscoring the need for hybrid approaches that blend automated translation with expert oversight.

2.4. General Text Embedding Models

Universal sentence embeddings have become foundational components in modern NLP, enabling transfer learning across a wide range of tasks such as retrieval, classification, semantic similarity, and clustering. This section surveys recent advances in the development of general-purpose embedding models and outlines how these advances informed the design and training strategy of our Turkish-specific models. The discussion is organized into two key components: (1) the construction and characteristics of large-scale weakly supervised sentence pair corpora, and (2) model architectures and training strategies used to derive universal text embeddings from these corpora.

2.4.1. Large-Scale Sentence Pair Corpora

The ability to train general-purpose embeddings hinges critically on the quality and scale of sentence pair corpora. Earlier approaches often relied on manually curated natural language inference (NLI) datasets or paraphrase corpora. However, recent models increasingly rely on large, noisy, but semantically aligned sentence pairs mined from web-scale corpora.

The C-MTP dataset introduced in C-MTEB (2) contains over 100 million Chinese sentence pairs mined from question answering (QA) forums, encyclopedic entries, and social media, utilizing retrieval heuristics and title-body matching. Similarly, the E5 framework (5) constructed a corpus from English and multilingual web data, including mC4, Wikipedia, QA sites (like StackExchange), and web documents. It used rule-based extraction strategies, such as aligning (title, body), (query, answer), or (instruction, response) to build over a billion sentence pairs with minimal human intervention.

The GTE model (6) follows a similar pipeline but emphasizes multilingual support. It integrates CCMatrix, Wikipedia translation pairs, and QA data across 100+ languages. Importantly, it introduces sentence deduplication, language filtering, and dynamic sampling to increase diversity and reduce noise.

Most recently, M3-Embedding (15) introduces a modular pipeline to extract dense, sparse, and multi-vector training pairs. This framework combines both explicit (instruction-response) and implicit (retrieval-based) alignments. It introduces automatic dataset cleaning via token-level filtering, sentence normalization, and low-resource augmentation strategies such as multilingual instruction expansion.

Inspired by these works, we curated a 62.5M Turkish sentence pair corpus from over 50 Hugging Face datasets. These sources span question answering, summarization, paraphrase mining, and instruction datasets, many of which are either natively

Turkish or translated using high-quality large language models (LLMs). To refine the quality, we trained several small-scale contrastive models and applied filtering based on cosine similarity thresholds. The final corpus of 34.2M sentence pairs integrates both symmetric (e.g., paraphrases) and asymmetric (e.g., question-answer, title-body) relations to support broad downstream generalization.

2.4.2. Training Universal Embedding Models

The learning of robust sentence embeddings from sentence pairs has been largely driven by the introduction of contrastive learning objectives. One of the most influential models in this domain is Sentence-BERT (SBERT) (16), which fine-tunes a Siamese BERT architecture using contrastive losses (triplet loss, cosine similarity, or classification heads) on NLI and STS datasets. SBERT showed that simple embedding pooling (e.g., mean pooling) over transformer outputs can yield general-purpose sentence representations.

Building on this, E5 (5) introduced the use of in-batch negatives via Multiple Negatives Ranking Loss (MNRL), where the loss is computed such that each positive sample is contrasted against all other negatives in the batch. To simulate large batch sizes, Cached-MNRL was introduced, which caches embeddings of previous batches to improve training dynamics. The multilingual variant mE5 (17) extended this approach to 100+ languages and applied a mix of multilingual instruction tuning and translation pair extraction.

The BAAI General Embedding (BGE) family (2) added instruction awareness by training models with a variety of prompts depending on downstream use cases. It used hybrid prompting for classification, retrieval, and STS and applied both contrastive and cross-entropy losses. BGE also introduced prompt-specific query encoders, enabling better alignment between embedding space structure and task expectations.

GTE (General Text Embedding) (6) emphasized robust multilingual alignment

and used knowledge distillation to transfer embedding structure from teacher models into smaller encoders. It employed a temperature-scaled MNRL loss and multilingual batch balancing during training. Finally, M3-Embedding (15) unified dense, sparse, and multi-vector embedding generation within a single encoder by mixing tasks during training. It used retrieval loss, classification loss, and cross-modal distillation simultaneously, demonstrating state-of-the-art results on BEIR and MTEB.

We adopt a similar two-stage design for Turkish: (1) contrastive pretraining using Cached-MNRL over the 34.2M sentence pair corpus, and (2) task-specific fine-tuning on TR-MTEB using appropriate loss functions: softmax cross-entropy for classification, cosine-based CoSENT for STS, and MNRL for retrieval. Our architecture is based on `bert-base-turkish-uncased`, making it computationally lightweight (110M parameters) yet effective across tasks. Despite limited resource availability, our model narrows the gap with multilingual counterparts and establishes strong Turkish-specific baselines.

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.

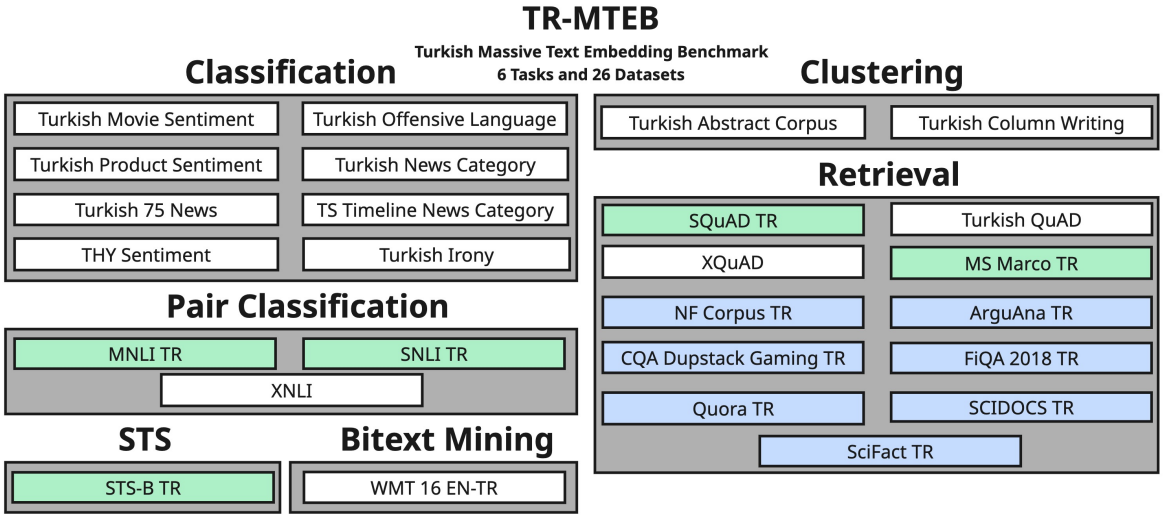


Figure 3.1. An overview of tasks and datasets included in TR-MTEB. Datasets in white background are original Turkish datasets. Datasets highlighted in green indicate those that were translated to Turkish prior to this study, while datasets marked in blue represent English datasets that were translated into Turkish as part of this work.

3.1. Turkish Massive Text Embedding Benchmark (TR-MTEB)

The development of the Turkish Massive Text Embedding Benchmark (TR-MTEB) was motivated by the severe under representation of Turkish in existing multilingual sentence embedding evaluation frameworks. At the time of writing, the multilingual MTEB benchmark[2] includes only three Turkish datasets, two for classification and one for retrieval, which limits its utility in assessing the true capabilities of Turkish-specific models. Moreover, evaluating a Turkish encoder on datasets originally constructed for other languages fails to account for Turkish-specific phenomena such as agglutination, rich morphology, and flexible word order.

To fill this gap, we introduce TR-MTEB, a task-diverse, Turkish-specific benchmark designed to comprehensively evaluate the performance of sentence embedding models across a variety of natural language understanding tasks. Our benchmark includes six core task types: classification, clustering, pair classification, bitext mining, semantic textual similarity (STS), and retrieval. Each task is adapted from the design principles of MTEB, but customized to align with the structure and semantics of Turkish-language data.

To ensure data quality and representativeness, we curated datasets from previously published academic sources. Only datasets with peer-reviewed documentation or prior use in published NLP research were selected. Where necessary, additional preprocessing or task reformulation was applied to adapt these datasets into a unified sentence embedding evaluation framework.

3.1.1. Tasks Overview

An overview of the benchmark's task structure is given in Figure 3.1. Below, we provide detailed descriptions of how each task is implemented in TR-MTEB:

[2]https://huggingface.co/spaces/mteb/leaderboard

3.1.1.1. Classification. In this task, sentence embeddings are used as input features to a downstream classifier trained on labeled data. Specifically, we use the training split of each dataset to train a logistic regression classifier on top of frozen embeddings, and evaluate it on the test split. The objective is to assess how well the embeddings separate semantically meaningful categories. For example, news articles might be classified into topics like sports, politics, or health. Metrics reported include Accuracy (primary), along with F1-score and Average Precision to reflect balanced and probabilistic performance.

3.1.1.2. Clustering. Clustering evaluates the intrinsic structure of the embedding space by measuring how well similar sentences group together without supervision. Given a test set with labeled categories (e.g., author labels, document topics), we compute sentence embeddings and cluster them using mini-batch k-means. The number of clusters is set equal to the number of gold labels. The primary evaluation metric is V-Measure, a harmonic mean of homogeneity (each cluster contains only members of a single class) and completeness (all members of a class are assigned to the same cluster). This task reflects whether semantically similar sentences are embedded near each other in high-dimensional space.

3.1.1.3. Pair Classification. Pair classification involves determining whether a given pair of sentences shares a specific semantic relationship, such as entailment, paraphrase, or duplication. For each pair, we compute the cosine similarity between their embeddings and apply a learned threshold to classify the pair as positive or negative. We train a logistic regression classifier or a simple thresholding mechanism on cosine scores from training pairs and evaluate on held-out sets. Metrics include Average Precision (primary), along with Accuracy, F1-score, Precision, and Recall. This task is particularly important for applications like semantic search, QA, and deduplication.

3.1.1.4. Bitext Mining. Bitext mining assesses cross-lingual alignment by identifying sentence pairs that are mutual translations. For each source sentence in Turkish, the

goal is to find its correct English counterpart from a large pool of candidates, and vice versa. Sentence embeddings are generated independently for each side of the corpus, and similarity scores are computed using cosine distance. For evaluation, we compute F1-score (primary), Precision, Recall, and Accuracy based on how many true translation pairs are correctly identified in top-n rankings. This task is critical for multilingual applications such as parallel corpus construction and machine translation.

3.1.1.5. Semantic Textual Similarity (STS). STS tasks measure how well sentence embeddings capture fine-grained semantic similarity on a continuous scale. Each sentence pair in the dataset is annotated with a human-judged similarity score, typically ranging from 0 (completely dissimilar) to 5 (semantically equivalent). For each pair, we compute the cosine similarity of their embeddings and evaluate how well this correlates with the gold-standard score using Spearman (primary) and Pearson correlation. This task directly evaluates the ability of embeddings to encode subtle meaning differences, and is a widely used intrinsic benchmark in sentence representation learning.

3.1.1.6. Retrieval. Retrieval simulates real-world information retrieval scenarios where a system must rank documents according to their relevance to a user query. In this task, each query is paired with a set of candidate documents, only some of which are relevant. We encode all queries and documents into embeddings and rank the candidates for each query based on cosine similarity. Evaluation metrics include nDCG10 (primary), Mean Reciprocal Rank (MRRk), Mean Average Precision (MAPk), Precisionk, and Recallk. This task is especially sensitive to the embedding model's ability to generalize across queries and long-form content.

Together, these tasks span both supervised and unsupervised settings, sentence-level and document-level contexts, and monolingual and cross-lingual scenarios. TR-MTEB thus serves as a comprehensive benchmark to assess generalization, semantic robustness, and domain transfer capabilities of Turkish embedding models. In the next subsections, we detail the collection of these datasets and the creation of additional

translated resources that further enhance the benchmark's coverage.

3.1.2. Collecting and Structuring TR-MTEB Datasets

To construct TR-MTEB, we systematically curated a diverse set of Turkish datasets spanning six core NLP tasks. We prioritized sources that have been used in peer-reviewed research and are available under open licenses. Each dataset was formatted to comply with the evaluation protocols defined in the original MTEB benchmark, ensuring consistency across task types. Below we describe the dataset selection and formatting process for each task category.

3.1.2.1. Classification. The classification task in TR-MTEB evaluates a model's ability to generate sentence embeddings that are linearly separable across predefined categories. To this end, we compiled a diverse set of eight Turkish datasets spanning multiple application domains, including sentiment analysis, news categorization, irony detection, and offensive language identification. Among these, six are newly collected Turkish datasets, while two are adopted from the original MTEB benchmark to ensure consistency and facilitate cross-lingual comparison.

For each dataset, we standardized the input format by pairing each sentence with its corresponding class label. In line with the MTEB evaluation protocol, we used embeddings of training examples to train a logistic regression classifier and reported Accuracy, F1-score, and Average Precision on the test set. Datasets with no predefined splits were used in a zero-shot setting, while others with train/validation/test splits enabled supervised training.

Below is a detailed overview of the datasets used for classification:

- THY Sentiment Classification: This dataset consists of 23.3k tweets sourced from Turkish Airlines' official social media accounts. Each tweet is labeled as positive or negative, capturing consumer sentiment in real-world customer service scenar-

ios (18).

- TS Timeline News Category: Derived from a large collection of Turkish news articles, this dataset includes 4k samples labeled across multiple categories such as politics, sports, economics, and health. It is designed to test the ability of embeddings to distinguish between semantically distinct news domains (19).
- 75 News Classification: A small-scale, balanced dataset composed of 75 Turkish news articles, equally divided into five topical classes (economics, magazine, health, politics, and sports). This resource is ideal for zero-shot evaluation due to its limited size (**?**).
- Turkish Irony Classification: This dataset contains 600 social media posts annotated for the presence or absence of irony. Given the subtle semantic cues involved in irony detection, it serves as a challenging benchmark for contextual understanding (20).
- Turkish Movie Sentiment Classification: Composed of over 10k Turkish movie reviews, this dataset is labeled for binary sentiment (positive/negative) and reflects user-generated text in the entertainment domain (21). Adopted from the original MTEB benchmark.
- Turkish News Category Classification: Includes 1,150 manually annotated news articles distributed over five predefined categories. Unlike the TS Timeline dataset, this corpus offers more concise examples and a clearer label distribution (22).
- Turkish Offensive Language Classification: This dataset includes over 34k Turkish tweets annotated as offensive or non-offensive. As part of a multilingual offensive speech corpus, it highlights the model's ability to capture social and pragmatic nuances (23).
- Turkish Product Sentiment Classification: This dataset contains product reviews collected from various Turkish e-commerce platforms, labeled for sentiment polarity. It complements the movie review dataset by offering domain-specific vocabulary and structure (21). Adopted from the original MTEB benchmark.

Each dataset was converted into a unified format and uploaded to Hugging Face to ensure reproducibility and ease of access. Table 3.1 summarizes these datasets,

including their train/validation/test splits and language origin. In the accompanying dataset table, we indicate the *Language Origin* of each resource using three categories: Native Turkish (originally created in Turkish), Pre-Translated (previously translated into Turkish in prior studies), and Translated in this study (newly translated from English as part of this work).

Table 3.1. Overview of classification datasets in TR-MTEB. All datasets are in Turkish and span diverse domains including social media, e-commerce, and news media.

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

This selection ensures that the classification task in TR-MTEB is representative of both formal and informal textual domains, enabling a comprehensive evaluation of the embedding models' generalizability and semantic discriminability.

3.1.2.2. Clustering. To assess the ability of sentence embeddings to capture underlying group structure in text, we include two Turkish datasets designed for document-level clustering. In this task, embeddings are computed for each document in the test set and then grouped using the mini-batch K-Means algorithm, with the number of clusters

set to the number of unique ground-truth labels. Evaluation is based on V-Measure, which combines homogeneity (purity of clusters) and completeness (coverage of classes) to provide a balanced metric for clustering performance.

The selected datasets span two distinct domains, academic writing and opinion journalism, allowing us to evaluate clustering performance in both formal and subjective genres:

- Turkish Abstract Corpus Clustering: This dataset contains 6,234 abstracts of Turkish academic papers, categorized by research discipline (24). Each document is labeled with a disciplinary category (e.g., engineering, law, or health sciences). The task here tests whether sentence embeddings can capture the topical structure and domain-specific terminology prevalent in academic discourse.
- Turkish Column Writing Clustering: This dataset includes 630 Turkish opinion articles written by 18 different authors, with 35 articles per author (25). Each article is labeled with its author and the task evaluates the model's ability to cluster texts based on stylistic and thematic consistency within an author's body of work.

Both datasets were used in their entirety for testing, without additional training or validation splits, consistent with the original MTEB protocol for clustering tasks. Table 3.2 provides a summary of dataset sizes, language origin, and Hugging Face access links.

Table 3.2. Overview of datasets used in the clustering task. Both datasets are used in a zero-shot setting with mini-batch K-Means and evaluated via V-Measure.

Dataset Name	**Train**	**Validation**	**Test**	**Language Origin**	**Hugging Face Link**
Turkish Abstract Corpus Clustering	-	-	6,234	Native Turkish	`https://huggingface.co/datasets/trmteb/ts_abstract_corpus_p2p`
Turkish Column Writing Clustering	-	-	630	Native Turkish	`https://huggingface.co/datasets/trmteb/630koseyazisi_p2p`

3.1.2.3. Pair Classification. The pair classification task evaluates whether sentence embeddings can effectively model the semantic relationship between two sentences, most commonly in the form of NLI. In TR-MTEB, this task is formulated as binary classification, distinguishing between `entailment` and `non-entailment` cases. For this purpose, we included three widely used datasets that have been machine-translated into Turkish or contain a Turkish subset, all of which are standard benchmarks for NLI in multilingual settings.

Each dataset is standardized into JSONL format with fields for `sentence1`, `sentence2`, and a binary `label`. Sentence embeddings are computed for each input pair and their cosine similarity is used as the predictive signal. A classifier or threshold is then trained on the training set and evaluated on the test set. The evaluation metrics include average precision, accuracy, F1-score, precision, and recall, in alignment with the MTEB framework.

The following datasets are included:

- MultiNLI-TR: This is a neural machine-translated version of the Multi-Genre Natural Language Inference (MultiNLI) dataset (26). It consists of sentence pairs across a variety of genres, making it a valuable resource for testing the robustness of sentence embeddings across different discourse domains. In our setup, we convert the original three-way labels (entailment, contradiction, neutral) into a binary format by labeling only `entailment` as positive and all others as negative.
- SNLI-TR: A Turkish translation of the Stanford Natural Language Inference (SNLI) dataset (26), originally designed for image captioning-based premise-hypothesis pairs. Like MultiNLI, the three-way NLI labels are binarized for our task. Its more controlled and homogeneous style complements MultiNLI's genre diversity.
- XNLI (Turkish Split): XNLI is a cross-lingual NLI benchmark that includes professional and machine-translated versions of dev and test sets in multiple languages (27). We use the Turkish split, which allows us to evaluate model perfor-

mance under cross-lingual transfer settings. In TR-MTEB, we follow the same binarization strategy for label consistency.

Table 3.3 summarizes the size and provenance of each dataset, along with their Hugging Face repository links for reproducibility.

Table 3.3. Overview of datasets used in the pair classification task. Labels were binarized to distinguish between `entailment` and `non-entailment`.

Dataset Name	**Train**	**Validation**	**Test**	**Language Origin**	**Hugging Face Link**
MultiNLI-TR	393k	10k	10k	Pre-Translated	`https://huggingface.co/datasets/trmteb/multinli_tr`
SNLI-TR	550k	10k	10k	Pre-Translated	`https://huggingface.co/datasets/trmteb/snli_tr`
XNLI (TR)	393k	2.49k	5.01k	Pre-Translated	`https://huggingface.co/datasets/mteb/xnli`

These datasets collectively provide a robust foundation for evaluating how well sentence embeddings capture semantic relationships in Turkish, especially under translation and genre variation.

3.1.2.4. Semantic Textual Similarity (STS). The Semantic Textual Similarity (STS) task evaluates how well sentence embeddings capture the degree of semantic relatedness between sentence pairs. In TR-MTEB, this task is formulated as a regression problem, where models are expected to produce similarity scores that correlate with human-annotated judgments on a continuous scale.

To benchmark this task, we employed the Turkish version of the STS Benchmark (28) (STSb TR), which is a machine-translated version of the original English STS-B dataset. The translation was performed using the Google Cloud Translation API, and no post-editing or human correction was applied. Despite the lack of human refinement, the dataset is retained due to its wide adoption in sentence similarity evaluation and the value it offers in assessing generalization to Turkish.

Each sample in the dataset consists of a sentence pair `(sentence1, sentence2)` and a corresponding similarity score ranging from 0 (no semantic similarity) to 5 (complete semantic equivalence). During evaluation, embeddings are computed for each sentence and their cosine similarity is compared to the gold-standard similarity scores using Spearman and Pearson correlation metrics. This setup aligns with the standard MTEB protocol for the STS task.

The dataset was not used for supervised fine-tuning due to the presence of translation artifacts and noise introduced during the automatic translation process. However, its inclusion as an evaluation-only benchmark allows us to measure the model's zero-shot performance on fine-grained semantic comparison. Table 3.4 summarizes the structure and provenance of the dataset.

Table 3.4. Overview of the dataset used in the semantic textual similarity task. Cosine similarity between sentence embeddings is evaluated using Spearman and Pearson correlations against the annotated similarity scores.

Dataset Name	**Train**	**Validation**	**Test**	**Language Origin**	**Hugging Face Link**
STSb TR	5.75k	1.5k	1.38k	Pre-Translated	`https://huggingface.co/datasets/trmteb/stsb-tr`

This task plays a crucial role in gauging the ability of Turkish embedding models to capture nuanced semantic relationships and supports the broader goal of enabling high-quality sentence representations for low-resource languages.

3.1.2.5. Bitext Mining. The Bitext Mining task in TR-MTEB evaluates the ability of sentence embedding models to align semantically equivalent sentences across languages, in this case, Turkish and English. This task is particularly relevant for applications such as machine translation, cross-lingual retrieval, and multilingual pretraining.

For this purpose, we include the English–Turkish subset of the WMT16 benchmark (**?**), which consists of parallel sentences translated by professional translators as part of the WMT16 shared translation task. Each instance in the dataset is a sen-

tence pair (sentence_en, sentence_tr) that are mutual translations of one another. We adopted the MTEB-compatible format where English and Turkish sentences are presented as separate corpora, each with its own unique sentence ID, and the task is framed as a retrieval problem from one language to the other.

During evaluation, each sentence in the source language (e.g., English) is embedded independently using the model under evaluation, and ranked against all sentences in the target language (e.g., Turkish) based on cosine similarity. The sentence from the target language with the highest similarity score is considered as the predicted translation. Performance is measured using the F1-score as the primary metric, with additional metrics including Precision, Recall, and Accuracy@1.

This task is particularly challenging for monolingual models, as they are not trained with any cross-lingual supervision and must generalize alignment solely based on Turkish data. As such, bitext mining serves as a strong indicator of cross-lingual generalization capacity. Table 3.5 summarizes the key properties of the dataset used in this task.

Table 3.5. Overview of the dataset used in the bitext mining task. The dataset contains professionally translated English–Turkish sentence pairs and is used to assess cross-lingual semantic alignment capabilities of embedding models.

Dataset Name	**Train**	**Validation**	**Test**	**Language Origin**	**Hugging Face Link**
WMT16 EN-TR	206k	1k	3k	Native Turkish	https://huggingface.co/datasets/trmteb/wmt16_en_tr

This setup provides a rigorous evaluation of multilingual alignment and is particularly useful for contrasting the performance of monolingual Turkish models with multilingual baselines trained on cross-lingual corpora.

3.1.2.6. Retrieval. The retrieval task in TR-MTEB is designed to evaluate the capability of sentence embeddings to rank relevant documents given a query. This task covers both native Turkish and machine-translated datasets across a wide range of

domains, including open-domain question answering, fact verification, scientific literature retrieval, financial queries, and community-based QA. Models are required to embed queries and documents independently, and cosine similarity is used to rank the documents for each query.

Each dataset adheres to the standardized MTEB retrieval format with three core components:

- `corpus.jsonl`: A set of documents/passages indexed by unique IDs.
- `queries.jsonl`: A collection of queries with corresponding unique IDs.
- `qrels.tsv`: Relevance judgments mapping query IDs to relevant document IDs.

We include 11 datasets covering both general-purpose and domain-specific retrieval settings. The datasets marked as "Translated in this study" (Table 3.6) were created via our LLM-based translation pipeline and are explained in detail in Section 3.1.3.

- SQuAD TR: A machine-translated version of the SQuAD v2.0 dataset (29), adapted for retrieval. Queries and answers were translated using Amazon Translate. We manually restructured this QA dataset into a retrieval setting.
- TQuAD: A native Turkish QA dataset focused on the history of Turkish and Islamic science (30). Originally developed for a national artificial intelligence competition, the dataset includes factual queries grounded in Turkish educational content.
- XQuAD: The Turkish subset of the multilingual XQuAD QA benchmark (31). We converted it to a retrieval format by aligning context passages and questions.
- MSMARCO TR: A large-scale Turkish machine-translated version of the MS-MARCO Passage Ranking dataset (32). This dataset is adapted to the MTEB format and represents a standard benchmark for retrieval at scale.
- NFCorpus TR: A medical information retrieval corpus (33), translated into Turkish in this study. It includes health-related consumer queries and professionally

curated document passages.

- ArguAna TR: This dataset, originally designed for argument retrieval (33), was translated into Turkish using an LLM-based pipeline. It contains argument-focused questions and relevant excerpts.
- CQA Dupstack Gaming TR: A Turkish translation of the "Gaming" subset of CQADupStack (34). It consists of community-driven QA pairs from Stack Exchange, targeting duplicate question detection.
- FiQA 2018 TR: A Turkish version of the financial QA benchmark (4), featuring questions related to financial topics and associated snippets.
- Quora TR: Translated from the Quora question duplication dataset (35). It assesses a model's ability to retrieve semantically equivalent questions across paraphrastic expressions.
- SCIDOCS TR: A Turkish-translated data set for retrieving scientific literature at the document level (36). Queries correspond to paper titles or abstracts, while the corpus consists of scientific articles.
- SciFact TR: A benchmark for scientific fact verification, adapted into Turkish (37). Queries are scientific claims and documents are abstracts from scientific literature, formatted for retrieval.

The datasets exhibit a wide range of corpus sizes and query counts, providing a robust evaluation of embedding models across different retrieval scenarios. Table 3.6 summarizes the dataset properties, including corpus/query sizes, splits (where available), language origin, and Hugging Face links.

This diverse set of retrieval datasets ensures robust evaluation across a variety of linguistic phenomena and content domains in Turkish.

3.1.3. Translation of Retrieval Datasets from BEIR Benchmark

To extend TR-MTEB's coverage of retrieval tasks, we translated seven diverse datasets from the BEIR benchmark into Turkish. These datasets cover a range of

Table 3.6. Overview of datasets used in the retrieval task. Corpus and query sizes are provided along with data split availability, translation status, and Hugging Face links.

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`

domains, including open-domain question answering, fact verification, scientific and financial document retrieval, and community-driven QA platforms. This translation effort follows recent trends in multilingual benchmark development (e.g., PL-MTEB), where translated datasets are used to enrich task diversity in low-resource languages.

3.1.3.1. LLM-based Translation Strategy. We employed an LLM-based pipeline to ensure translation quality. Model selection was guided by the Cetvel Benchmark (**?**), a comprehensive suite for evaluating Turkish-language performance in LLMs. In the machine translation category of Cetvel, the 35B version of Cohere's `aya-expanse` (38) model ranked highest, but due to resource constraints, we used the 8B version, which provided comparable output quality with improved inference speed. Before finalizing the translation setup, we conducted a preliminary evaluation of several open-source and proprietary LLMs on a representative translation sample. Among the open-source models, `aya-expanse-8B` consistently delivered the highest quality outputs. While cer-

tain proprietary models exhibited superior translation fidelity, their API-based access posed significant cost constraints for large-scale translation, rendering them impractical for this study. To ensure consistency and semantic fidelity in translations, we used a structured prompt to instruct the LLM during the generation process. The prompt emphasized preservation of meaning, stylistic fluency, and proper handling of technical content. Figure 3.2 gives the exact instruction issued to the model for each translation call.

```
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:
{original_text}

# Translated Text:
```

Figure 3.2. Prompt for Translating BEIR Dataset into Turkish

3.1.3.2. Translation Evaluation: LLM-as-a-Judge. To validate the translation quality, we developed an automatic evaluation mechanism, referred to as *LLM-as-a-Judge*, in which a large language model assessed the quality of machine-generated Turkish translations through a binary `PASS`/`FAIL` decision. The evaluation was carried out using the same `aya-expanse-8B` model employed in the translation stage and the evaluation

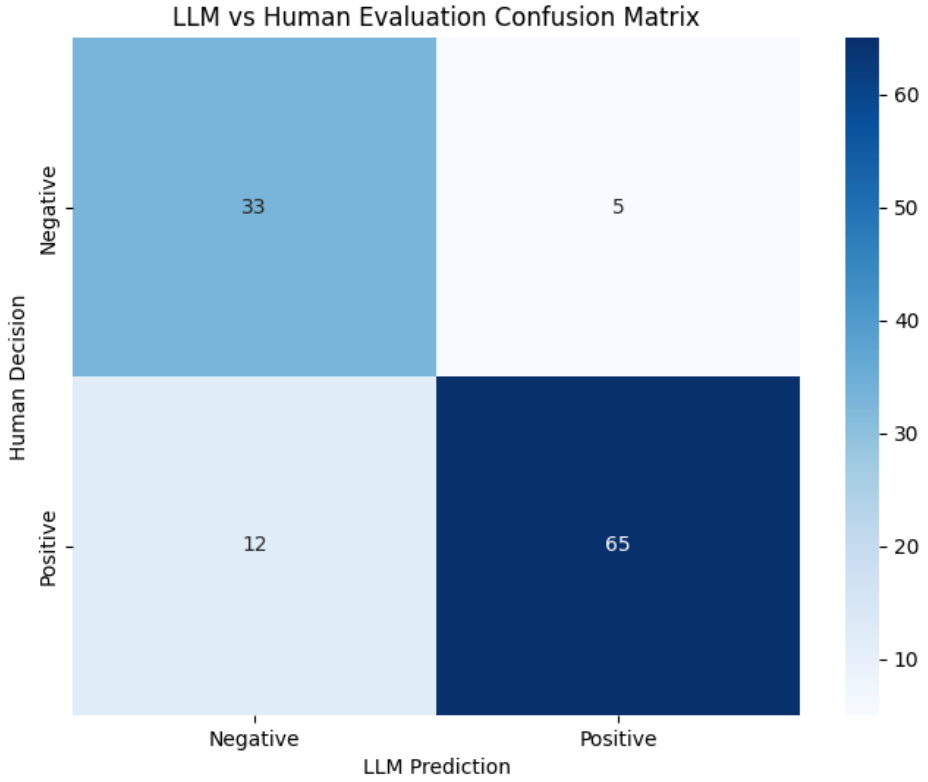


Figure 3.3. Confusion matrix comparing human annotations and LLM predictions for translation quality evaluation.

was guided by a structured evaluation prompt (Listing 3.4) that outlined key quality dimensions including semantic accuracy, fluency, consistency, and cultural appropriateness. To calibrate the prompt, we manually labeled 115 examples drawn uniformly from all seven translated datasets. Each example was independently annotated by human raters based on semantic fidelity and naturalness. The LLM prompt was then refined iteratively to align its predictions with human judgment. The final version demonstrated high agreement, achieving 85.2% exact match, 92.9% precision, 84.4% recall, and an F1-score of 88.4%, as illustrated in Figure 3.3.

```
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
}
```

Figure 3.4. Prompt template for LLM-as-a-Judge translation evaluation

3.1.3.3. Post-Editing Pipeline. After validating the evaluator, we applied it across all translated examples. Translations flagged as `FAIL` were manually post-edited or re-generated using stronger models via Google AI Studio[3] . This hybrid pipeline allowed us to ensure that the final datasets retained task integrity and translation accuracy, despite being automatically generated.

After post-editing was completed, we performed an additional round of manual validation to assess the final translation quality. Specifically, we randomly selected 5 queries and 5 corpus entries from each of the 7 translated datasets (70 samples in total) and manually scored their translation quality on a 0–10 scale. The resulting average score was **92.71%**, indicating that while most translations were of high quality, a few errors persisted even after the LLM-as-a-Judge filtering. Upon further inspection, common issues included: (1) the LLM occasionally answering a query instead of translating it literally, despite instructions in the prompt not to do so; (2) mild readability issues due to syntactically awkward or inverted sentence structures in longer translations; (3) incorrect handling of proper noun suffixes in Turkish (e.g., mistranslating Ankara'nın as Ankara'yı); and (4) minor character-level mistakes in the translation of scientific terminology. These findings suggest that while LLM-based translation pipelines are highly effective, careful human verification remains critical for benchmark-quality datasets.

3.1.3.4. Downstream Use. The final Turkish versions of these retrieval datasets, `FiQA`, `NFCorpus`, `CQADupstack-Gaming`, `ArguAna`, `SciDocs`, `SciFact`, and `Quora`, were integrated into the TR-MTEB suite for evaluation. They were standardized into MTEB-compatible formats (`corpus.jsonl`, `queries.jsonl`, `qrels.tsv`) and can be used to benchmark Turkish sentence embeddings under realistic retrieval conditions. Each dataset is released under a research policy ensuring broad accessibility while respecting the source conditions of the original English corpora.

[3]https://aistudio.google.com/

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. In this context, weakly supervised refers to sentence pairs that were not manually labeled but are assumed to be semantically related based on structural heuristics (e.g., question-answer, title-content) or derived from high-quality sources. These pairs offer a scalable alternative to human annotation, enabling the training of sentence embedding models without the need for costly supervision.

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 generated resources using LLMs. A detailed list of the dataset sources, their schema types, and filtering statistics is provided in Table 3.7.

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

Table 3.7. 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**	**Ref. (if any)**	**Hugging Face Link**
WikiRAG-TR	Q&A	Synthetically Generated	18k	9.73k	-	https://huggingface.co/datasets/Metin/WikiRAG-TR
wikisource	Title&Text	Native Turkish	7.22k	4.07k	(39)	https://huggingface.co/datasets/wikimedia/wikisource
mqa	Q&A	Native Turkish	6.66M	3.79M	(40)	https://huggingface.co/datasets/clips/mqa
mfaq	Q&A	Native Turkish	102k	61.1k	(40)	https://huggingface.co/datasets/clips/mfaq
tapaco	Paraphrase	Native Turkish	183k	168k	(41)	https://huggingface.co/datasets/community-datasets/tapaco
Turkish Medical Reasoning	Q&A with Content	Translated	7.21k	6.96k	-	https://huggingface.co/datasets/ituperceptron/turkish_medical_reasoning
wiki_lingua	Title & Content	Native Turkish	4.5k	2.73k	(42)	https://huggingface.co/datasets/esdurmus/wiki_lingua
oasst1 Pairwise RLHF Reward	Q&A	Instruction	9	5	-	https://huggingface.co/datasets/tasksource/oasst1_pairwise_rlhf_reward

Table 3.7. 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**	**Ref. (if any)**	**Hugging Face Link**
Wikipedia TR Summarization	Summarization	Native Turkish	125k	125k	(43)	https://huggingface.co/datasets/musabg/wikipedia-tr-summarization
Turkish Instructions	Instruction	Translated	51.9k	42k	-	https://huggingface.co/datasets/SoAp9035/turkish_instructions
Wikinews Multilingual	Title & Content	Native Turkish	363	252	(44)	https://huggingface.co/datasets/Fumika/Wikinews-multilingual
OpenOrca TR	Instruction	Translated	798k	649k	(45)	https://huggingface.co/datasets/ucekmez/OpenOrca-tr
ARC TR	Q&A	Translated	31.1k	4.41k	(46)	https://huggingface.co/datasets/malhajar/arc-tr
OO GPT4 Filtered TR	Q&A	Translated	101k	85.8k	-	https://huggingface.co/datasets/umarigan/oo-gpt4-filtered-tr
babel briefings	Q&A with Context	Native Turkish	341k	260,1k	(47)	https://huggingface.co/datasets/felixludos/babel-briefings
flickr8k TR	Paraphrase	Synthetically Generated	8k	6.48k	(48)	https://huggingface.co/datasets/atasoglu/flickr8k-turkish
SE dataset	Instruction	Native Turkish	877	765	-	https://huggingface.co/datasets/salihturkoglu/se_data_set

Table 3.7. 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**	**Ref. (if any)**	**Hugging Face Link**
Turkish Youtube Comments	Comment & Reply	Native Turkish	8.91k	3.49k	-	https://huggingface.co/datasets/yusiqo/Turkish-Youtube-Comments
Turkish Medical QA	Q&A	Native Turkish	188k	75.1k	(49)	https://huggingface.co/datasets/alibayram/doktorsitesi
Onedio Haberler	Title & Content	Native Turkish	66.7k	59.5k	(50)	https://huggingface.co/datasets/alibayram/onedio_haberler
CC News	Title & Content	Native Turkish	3.72M	3.25M	-	https://huggingface.co/datasets/stanford-oval/ccnews
Neural News Benchmark	Title & Content	Native Turkish	3.21k	1.92k	(51)	https://huggingface.co/datasets/tum-nlp/neural-news-benchmark
gsm8k TR	Q&A	Translated	8.76k	8.74k	(52)	https://huggingface.co/datasets/bezir/gsm8k-tr
Instruc Turca	Instruction	Synthetically Generated	363	252	(53)	https://huggingface.co/datasets/turkish-nlp-suite/InstrucTurca
Turkish Law Chatbot	Q&A	Native Turkish	14.9k	12.5k	-	https://huggingface.co/datasets/Renicames/turkish-law-chatbot

Table 3.7. 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**	**Ref. (if any)**	**Hugging Face Link**
Wikipedia 2024-06 bge-m3	Title & Content	Native Turkish	2.05M	1.25M	-	https://huggingface.co/datasets/Upstash/wikipedia-2024-06-bge-m3
alpaca TR	Instruction	Translated	45.3k	27.7k	-	https://huggingface.co/datasets/BrewInteractive/alpaca-tr
Turkish LLM Finetune Dataset	Q&A with Context	Native Turkish	193.6k	150.9k	-	https://huggingface.co/datasets/barathanasln/turkish_llm_finetune_dataset_4_topics
TRT Data Warrriors Dataset	Q&A	Native Turkish	2k	1.97k	-	https://huggingface.co/datasets/NusretOzates/TRTDataWarrriorsDataset
Multilingual Reward Bench	Instruction	Native Turkish	2.87k	2.59k	(54)	https://huggingface.co/datasets/CohereLabsCommunity/multilingual-reward-bench
TDK Sozluk Turkish	Paraphrase	Native Turkish	133k	72.8k	-	https://huggingface.co/datasets/Ba2han/TDK_Sozluk-Turkish
turkish exam instructions	Instruction	Native Turkish	41.4k	39.9k	(55)	https://huggingface.co/datasets/bezir/turkish_exam_instructions

Table 3.7. 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**	**Ref. (if any)**	**Hugging Face Link**
Wikipedia Turkish QA Chat	Q&A	Native Turkish	1.6M	784k	-	https://huggingface.co/datasets/Quardo/wikipedia-turkish-qa-chattemplate
News TR	Title & Content	Native Turkish	1.53M	1.36M	-	https://huggingface.co/datasets/habanoz/news-tr-1.8M
Med Turk QuAD	Q&A with Context	Native Turkish	24.6k	7.82k	(56)	https://huggingface.co/datasets/incidelen/MedTurkQuAD
Legal NLI TR	Paraphrase	Native Turkish	203k	192k	-	https://huggingface.co/datasets/Turkish-NLI/legal_nli_TR_V1
turoqa small	Instruction	Native Turkish	3k	3k	-	https://huggingface.co/datasets/SoAp9035/turoqa-small
Turkish Wikipedia QA	Q&A	Native Turkish	4.01M	2.88M	-	https://huggingface.co/datasets/hcsolakoglu/turkish-wikipedia-qa-4-million
Thinking Data Turkish	Instruction	Translated	209k	202k	-	https://huggingface.co/datasets/erythropygia/ThinkingData-200K-Turkish
Multilingual NLI	Paraphrase	Translated	33.82k	26.62k	(?)	https://huggingface.co/datasets/MoritzLaurer/multilingual-NLI-26lang-2mil7

Table 3.7. 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**	**Ref. (if any)**	**Hugging Face Link**
Seahorse Summarization	Text & Summary	Synthetically Generated	696.7k	310k	(57)	https://huggingface.co/datasets/tasksource/seahorse_summarization_evaluation
lr-sum	Text & Summary with Title	Synthetically Generated	107.4k	87.6k	(58)	https://huggingface.co/datasets/bltlab/lr-sum
LlamaTurk Instruction Set	Instruction	Translated	52k	41.6k	-	https://huggingface.co/datasets/metunlp/LlamaTurk-Instruction-Set
xMIND Large	Text & Summary	Native Turkish	279k	202k	(59)	https://huggingface.co/datasets/aiana94/xMINDlarge
Patient-Doctor QA	Q&A	Translated	234.9k	189.1k	(60)	https://huggingface.co/datasets/kayrab/patient-doctor-qa-tr-19583
zynp_ai teknofest	Q&A	Native Turkish	35M	14M	(61)	https://huggingface.co/datasets/sekerlipencere/zynpdata-zynp_ai-teknofest
Turkish Headline Dataset	Text & Summary	Native Turkish	14.6k	12.2k	-	https://huggingface.co/datasets/nuilbg/turkish_headline_dataset_new
MLSUM	Summarization	Native Turkish	822k	721k	(62)	https://huggingface.co/datasets/reciTAL/mlsum

Table 3.7. 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**	**Ref. (if any)**	**Hugging Face Link**
TRSUM	Summarization	Native Turkish	921k	827k	(63)	-
Turkish Labeled Text Corpus	Title & Summary	Native Turkish	7.9k	6.61k	(64)	-

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.

- Instruction datasets: These typically contain prompts or instructions paired with model-generated completions or human responses. We treated (instruction, response) pairs as sentence pairs, sometimes augmenting with (instruction, concatenated context) pairs when additional information was provided.
- Summarization datasets: In datasets where a summary was provided for a longer document or passage, we constructed pairs like (summary, document), (summary, title), and (title, document), depending on the available fields. These pairs often reflect asymmetric relationships, capturing content condensation.
- Text-Summary with Title datasets: A richer variant of summarization datasets that contain a title, full article content, and a summary. We leveraged this structure to extract multiple combinations: (title, content), (title, summary), (summary, content), enabling multiple semantic relations.
- Q&A with Context datasets: These contain a question, an answer, and an associated passage or context from which the answer was derived. We generated all pairwise combinations between these components, such as (question, answer), (question, context), and (answer, context).
- Comment-Reply datasets: In user-generated content platforms such as social media or forums, comments and their replies often form informal Q&A structures. We extracted (comment, reply) pairs directly and sometimes used thread-level context to create (comment, thread-context) pairs.
- Instruction with Reward/Preference datasets: These datasets, often created for reinforcement learning from human feedback (RLHF), contain two responses to the same instruction, along with a label indicating the preferred one. We used the (instruction, preferred response) pair as the primary supervision and discarded the rejected variant to avoid noisy signals.
- Dialogue datasets: When available, conversational datasets were parsed to generate turn-based sentence pairs, such as (utterance_i, utterance_i+1). In longer conversations, we optionally included context windows to generate pairs like (utterance, dialogue context).

The resulting dataset includes a deliberate mix of symmetric and asymmetric sen-

tence 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. An overview of the distribution of sentence pairs across these dataset types is presented in Figure 3.5.

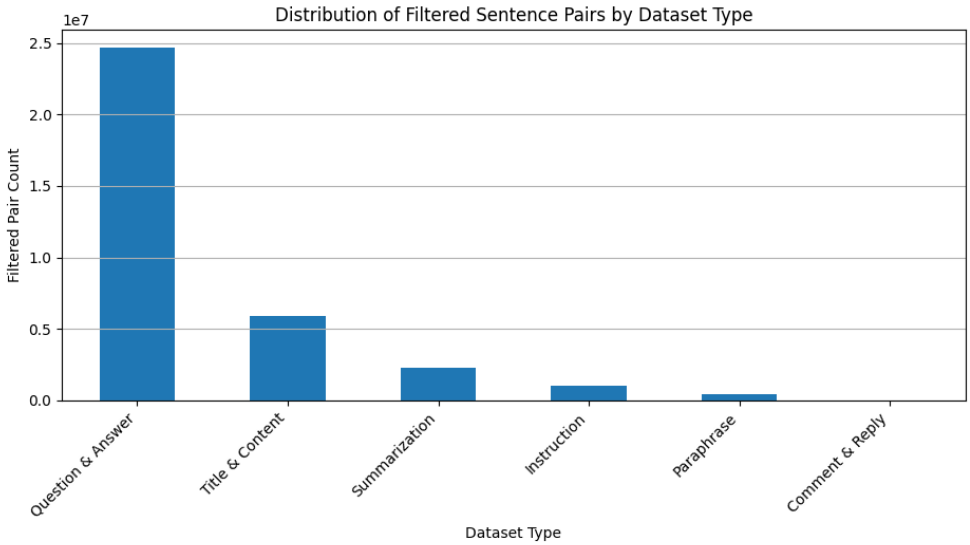


Figure 3.5. Distribution of filtered sentence pairs by dataset type. For visualization purposes, specific dataset types were grouped into broader categories such as *Question & Answer*, *Title & Content*, *Instruction*, etc. This aggregation simplifies presentation and enhances interpretability while reflecting the core structural formats used for weak supervision.

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.

Table 3.8. 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 4.2.

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

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 (62), (64), (63). 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 3.8) where *Mean (Task)* refers to the average score across datasets and *Mean (Task Type)* is averaged over task categories.

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 simi-

larity threshold of 0.4 as the most inclusive yet reasonably precise cut-off. 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 improved 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.

3.2.1. Training Turkish Sentence Embedding Models

To evaluate the effectiveness of our constructed corpus and to develop robust sentence embedding baselines for Turkish, we trained monolingual models using high-quality, weakly supervised sentence pairs. Our models are initialized from `bert-base-turkish-uncased`, a pretrained Turkish-specific BERT model developed by DBMDZ (7). Unlike multilingual encoders trained with extensive resources across dozens of languages, our approach aims to investigate whether Turkish-only contrastive learning on curated sentence pairs can produce competitive semantic representations.

3.2.1.1. Contrastive Pretraining with Cached-MNRL. The first stage involves unsupervised contrastive pretraining using the full set of 34.2 million filtered Turkish sentence pairs. For this, we employed the Cached Multiple Negatives Ranking Loss

[5]https://huggingface.co/datasets/trmteb/turkish_embedding_model_training_data
[6]https://huggingface.co/datasets/trmteb/cleaned_turkish_embedding_model_training_data_colab

(Cached-MNRL), which extends the conventional Multiple Negatives Ranking Loss (MNRL) by enabling effective large-batch training under constrained computational resources (65).

The standard MNRL formulation is as follows:

$$\mathcal{L}_{\text{MNRL}} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\exp(\cos(a_i, p_i)/\tau)}{\sum_{j=1}^{n}\exp(\cos(a_i, p_j)/\tau)} \tag{3.1}$$

where a_i and p_i represent the embeddings of the anchor and positive sentence in the i-th pair, respectively, and τ is a temperature hyperparameter. All p_j $(j \neq i)$ within the same batch act as implicit negatives. This formulation encourages the anchor to be closer to its positive match than to other unrelated positives within the batch.

Cached-MNRL builds on this by maintaining a memory bank (or cache) of previous positive embeddings, which effectively expands the negative pool without increasing GPU memory usage. By aggregating cached positives from prior iterations, we simulate a virtual batch size of up to 32,768, dramatically improving the contrastive signal during training.

3.2.1.2. Supervised Fine-Tuning. After unsupervised pretraining, we further refined the model using task-specific supervised datasets from TR-MTEB to align representations with evaluation targets. Fine-tuning was conducted on selected training splits for classification, pair classification, STS, and retrieval tasks.

Depending on the task format, we used the following objective functions:

- Classification: For single-sentence classification tasks (e.g., sentiment, topic detection), we used a standard cross-entropy loss applied on top of sentence embed-

dings passed through a linear classifier

$$\mathcal{L}_{\text{CLS}} = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) \tag{3.2}$$

where y_i is the probability of the true label and $\hat{y}_i$ is the probability of the predicted label.

- Semantic Similarity (STS): For sentence pair regression, we adopted the CoSENT (Cosine Similarity-based Entailment) loss (66)

$$\mathcal{L}_{\text{CoSENT}} = \log\left(\sum_{i,j;\, y_i > y_j} \exp\left(\alpha \cdot (\cos(\mathbf{u}_i, \mathbf{v}_i) - \cos(\mathbf{u}_j, \mathbf{v}_j))\right)\right) \tag{3.3}$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ denote the embeddings of sentence pairs, y_i is the gold similarity score, and α is a temperature scaling parameter. This loss encourages the model to assign higher cosine similarity to sentence pairs with greater semantic relatedness.

- Retrieval: For query-document retrieval tasks, we reused the MNRL loss with in-batch negatives, similar to pretraining.

3.2.1.3. Training Configuration. All training was conducted in Python using the `transformers` (67), `datasets` (68), and `sentence-transformers` (16) libraries. Pretraining was conducted on a single NVIDIA A100 GPU using Google Colab, with a total training time of approximately 80 GPU hours. Fine-tuning took an additional 2 GPU hours. We used a learning rate of 2e-5, batch size of 64 (effective size via cache: 32,768), and cosine learning rate scheduling with linear warmup.

3.2.1.4. Evaluation Pipeline. To fairly assess performance during fine-tuning, we used a modular evaluation setup with task-specific scoring scripts. For example, classification tasks were evaluated with accuracy and F1, STS with Spearman/Pearson correlation, and retrieval with nDCG10. Evaluation was performed on TR-MTEB validation splits, with best checkpoints selected using early stopping based on task-relevant met-

rics.

3.2.1.5. Our Models. We introduce two Turkish-specific sentence embedding models trained using the `bert-base-turkish-uncased` backbone. Both models are developed to explore the effect of domain-specific contrastive learning and supervised fine-tuning under limited-resource settings.

(i) `turkish-embedding-model`[7] : Contrastively pretrained on 34.2M filtered Turkish sentence pairs using Cached Multiple Negatives Ranking Loss (Cached-MNRL).
(ii) `turkish-embedding-model-fine-tuned`[8] : Fine-tuned version of the above model using labeled TR-MTEB training sets across classification, STS, retrieval, and pair classification tasks.

This two-stage pipeline, contrastive pretraining followed by supervised fine-tuning, enabled the construction of high-quality, Turkish-only embedding models that establish strong baselines on the TR-MTEB benchmark and pave the way for further low-resource embedding research.

[7]https://huggingface.co/trmteb/turkish-embedding-model
[8]https://huggingface.co/trmteb/turkish-embedding-model-fine-tuned

4. EXPERIMENTS AND RESULTS

4.1. Model Selection Strategy and Model Specifications

To provide a comprehensive evaluation across model architectures, training paradigms, and linguistic coverage, we selected a diverse set of sentence embedding models. Our goal was twofold: (1) to benchmark state-of-the-art multilingual and English-centric models against Turkish tasks; and (2) to understand how lightweight, monolingual Turkish models compare when trained on high-quality, task-relevant data.

The evaluated models span six general categories:

- Multilingual SOTA Models: These include the `multilingual-e5` family and `gte-multilingual-base`, which represent current state-of-the-art encoders in cross-lingual and multilingual settings.
- Instruction-Tuned Models: `multilingual-e5-large-instruct` is included to measure the effect of task-specific tuning objectives.
- Sentence-BERT Variants: Popular models from the Sentence Transformers library, such as `LaBSE`, `mpnet`, and `MiniLM`, serve as lightweight yet competitive multilingual and English-centric baselines.
- Monolingual Turkish Model: `bert-base-turkish-uncased`, a BERT model trained exclusively on Turkish corpora, is used both as a baseline and as the initialization point for our custom models.
- Static Word Embeddings: `fasttext-turkish` and `glove-turkish` represent pre-contextualized baselines, allowing us to contrast deep contextualized approaches with traditional embeddings.
- API-based Proprietary Model: `text-embedding-3-small` from OpenAI offers a high-performance proprietary baseline available via API.

As detailed in Section 3.2.1, we also release two Turkish-specific sentence embed-

ding models (highlighted in light gray in Table 4.1) trained using contrastive and supervised learning. These models are compact, accessible, and trained entirely on open-source Turkish data. All encoder-based models in this evaluation were used with a maximum input context length of 512 tokens, while the proprietary `text-embedding-3-small` model supports up to 8192 tokens. For static embedding baselines (e.g., FastText and GloVe), context length is not applicable due to their non-contextual nature. Input sequences longer than the model's maximum context were truncated from the end during preprocessing. Table 4.1 provides an overview of all models evaluated in this study, including their size, dimensionality, language coverage, reference source, and access link.

Table 4.1. Overview of sentence embedding models used in TR-MTEB experiments, extended with context length.

Model Name	**Size**	**Dim.**	**Lang.**	**Ctx. Len.**	**Ref.**	**Accessible URL**
multilingual-e5-large	560M	1024	Multi	512	(5)	`https://huggingface.co/intfloat/multilingual-e5-large`
multilingual-e5-large-instruct	560M	1024	Multi	512	(17)	`https://huggingface.co/intfloat/multilingual-e5-large-instruct`

Table 4.1. Overview of sentence embedding models used in TR-MTEB experiments, extended with context length (cont.).

Model Name	**Size**	**Dim.**	**Lang.**	**Ctx. Len.**	**Ref.**	**Accessible URL**
gte-multilingual-base	305M	1024	Multi	512	(6)	https://huggingface.co/Alibaba-NLP/gte-multilingual-base
multilingual-e5-base	278M	768	Multi	512	(5)	https://huggingface.co/intfloat/multilingual-e5-base
multilingual-e5-small	118M	384	Multi	512	(5)	https://huggingface.co/intfloat/multilingual-e5-small
LaBSE	471M	768	Multi	512	(69)	https://huggingface.co/sentence-transformers/LaBSE

Table 4.1. Overview of sentence embedding models used in TR-MTEB experiments, extended with context length (cont.).

Model Name	**Size**	**Dim.**	**Lang.**	**Ctx. Len.**	**Ref.**	**Accessible URL**
paraphrase-multilingual-mpnet-base-v2	278M	768	Multi	512	(16)	https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2
paraphrase-multilingual-MiniLM-L12-v2	118M	768	Multi	512	(16)	https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
all-MiniLM-L12-v2	33M	384	English	512	(16)	https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2
all-MiniLM-L6-v2	22M	384	English	512	(16)	https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Table 4.1. Overview of sentence embedding models used in TR-MTEB experiments, extended with context length (cont.).

Model Name	**Size**	**Dim.**	**Lang.**	**Ctx. Len.**	**Ref.**	**Accessible URL**
all-mpnet-base-v2	109M	768	English	512	(16)	https://huggingface.co/sentence-transformers/all-mpnet-base-v2
bert-base-turkish-uncased	110M	768	Turkish	512	(7)	https://huggingface.co/dbmdz/bert-base-turkish-uncased
fasttext-turkish	-	300	Turkish	–	(70)	https://fasttext.cc/docs/en/crawl-vectors.html
glove-turkish	-	–	Turkish	–	(71)	https://github.com/Turkish-Word-Embeddings/Word-Embeddings-Repository-for-Turkish

Table 4.1. Overview of sentence embedding models used in TR-MTEB experiments, extended with context length (cont.).

Model Name	**Size**	**Dim.**	**Lang.**	**Ctx. Len.**	**Ref.**	**Accessible URL**
text-embedding-3-small	N/A	768	English (API)	8192	(?)	https://platform.openai.com/docs/models/text-embedding-3-small
turkish-embedding-model (ours)	110M	768	Turkish	512	This work	https://huggingface.co/trmteb/turkish-embedding-model
turkish-embedding-model-fine-tuned (ours)	110M	768	Turkish	512	This work	https://huggingface.co/trmteb/turkish-embedding-model-fine-tuned

4.2. Performance Analysis and Evaluation Insights

To evaluate sentence embedding models for Turkish, we benchmarked a diverse set of multilingual and monolingual encoders on the TR-MTEB benchmark. Table 4.2 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 the original metrics while enabling standardized assessment on TR-MTEB.

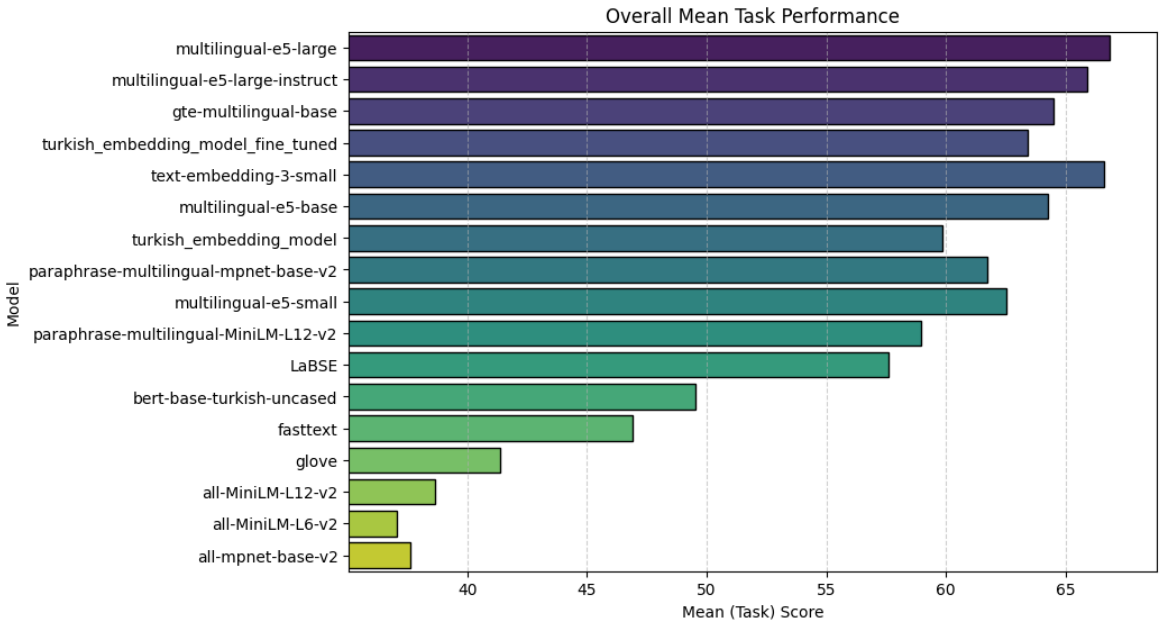


Figure 4.1. Overall Mean(Task) performance of sentence embedding models on TR-MTEB.

Table 4.2. TR-MTEB benchmark results across embedding models. Bitext: Bitext Mining, Class.: Classification, Clust.: Clustering, Pair Cl.: Pair Classification, Retr.: Retrieval, STS: Semantic Textual Similarity. Results ordered by Mean (Task) scores.

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

Table 4.2. TR-MTEB benchmark results across embedding models. Bitext: Bitext Mining, Class.: Classification, Clust.: Clustering, Pair Cl.: Pair Classification, Retr.: Retrieval, STS: Semantic Textual Similarity. Results ordered by Mean (Task) scores (cont.).

Model	**Dim**	**Mean (Task)**	**Mean (Type)**	**Bitext**	**Class.**	**Clust.**	**Pair Cl.**	**Retr.**	**STS**
text-embedding-3-small	768	66.61	69.52	91.55	69.51	62.08	58.13	64.99	70.85
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
turkish-embedding-model-fine-tuned	768	63.42	70.25	86.88	70.93	61.49	71.03	52.75	78.44
multilingual-e5-small	384	62.53	69.55	97.46	65.64	59.14	62.34	56.53	76.18
paraphrase-multilingual-mpnet-base-v2	768	61.71	71.05	96.06	69.06	57.52	72.22	49.27	82.18

Table 4.2. TR-MTEB benchmark results across embedding models. Bitext: Bitext Mining, Class.: Classification, Clust.: Clustering, Pair Cl.: Pair Classification, Retr.: Retrieval, STS: Semantic Textual Similarity. Results ordered by Mean (Task) scores (cont.).

Model	**Dim**	**Mean (Task)**	**Mean (Type)**	**Bitext**	**Class.**	**Clust.**	**Pair Cl.**	**Retr.**	**STS**
turkish-embedding-model	768	59.84	59.49	37.70	70.68	63.45	57.65	52.54	74.93
paraphrase-multilingual-MiniLM-L12-v2	768	58.94	68.44	93.67	65.76	56.64	69.56	46.57	78.41
LaBSE	768	57.60	66.60	99.53	65.60	58.40	57.83	46.47	71.75
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-mpnet-base-v2	768	37.62	38.06	5.07	54.43	46.63	49.79	22.26	50.21

Table 4.2. TR-MTEB benchmark results across embedding models. Bitext: Bitext Mining, Class.: Classification, Clust.: Clustering, Pair Cl.: Pair Classification, Retr.: Retrieval, STS: Semantic Textual Similarity. Results ordered by Mean (Task) scores (cont.).

Model	**Dim**	**Mean (Task)**	**Mean (Type)**	**Bitext**	**Class.**	**Clust.**	**Pair Cl.**	**Retr.**	**STS**
all-MiniLM-L6-v2	384	37.06	37.02	6.78	52.46	40.57	50.52	23.25	48.55

This section provides an in-depth technical analysis of the TR-MTEB benchmark results, structured around task-specific performance, training strategies, model architecture effects, and language-specific challenges. Insights are grounded in Table 4.2 and visualized through Figures 4.1 and 4.2. Figure 4.1 offers a comparative ranking of models based on their Mean(Task) scores, showcasing how our Turkish-specific models approach or match the performance of larger multilingual encoders despite their smaller size and training scope. Detailed results of the task-specific evaluation are provided in Appendix A, where each of the six core TR-MTEB task categories, Classification, Clustering, Bitext Mining, Pair Classification, Semantic Textual Similarity (STS) and Retrieval, are reported in tabular format. These tables break down the model performance per dataset, offering a granular view beyond the aggregate scores in Table 4.2.

4.2.1. Global Trends in Mean(Task) Performance

Multilingual models with large-scale pretraining dominate the overall rankings. `multilingual-e5-large` (66.82) and `text-embedding-3-small` (66.61) achieve the best Mean(Task) scores due to their massive training data, model capacity (up to

560M parameters), and instruction-tuning. These encoders are trained across a variety of tasks and languages, enabling them to generalize broadly across TR-MTEB.

In contrast, our monolingual `turkish-embedding-model-fine-tuned`, trained from scratch using only Turkish data, achieves 63.42 Mean(Task), outperforming several multilingual baselines despite having just 110M parameters and orders-of-magnitude less training data.

4.2.2. Performance Across Task Types

4.2.2.1. Classification Tasks. Classification shows relatively stable performance across the models. Our fine-tuned model reaches 70.93, surpassing most baselines and coming close to `multilingual-e5-base`. This task benefits from rich syntactic cues in Turkish and rewards the models that encode domain-specific structures—demonstrating the advantage of monolingual pretraining. For complete dataset-level results across classification tasks, see Table A.1 in Appendix A.

4.2.2.2. Clustering Tasks. With 61.49 on clustering tasks, our fine-tuned model again competes closely with multilingual alternatives. This task relies heavily on semantic grouping, and the success of our model reflects the effectiveness of weakly supervised contrastive training in capturing distributional semantics at the corpus level. Clustering results across academic abstracts and opinion columns are likewise reported in Table A.2.

4.2.2.3. Bitext Mining Task. This task poses the greatest challenge for monolingual models. Our contrastively trained model scores only 37.7, while the fine-tuned version reaches 86.88. Yet, both are outperformed by cross-lingual models like `LaBSE` and `e5-large`, each above 99. This result reflects the lack of English exposure and bilingual alignment in our training regime. Model performance on the WMT16 EN–TR bitext mining dataset is reported in Table A.3, reinforcing the gap between monolingual and

multilingual encoders in cross-lingual settings.

4.2.2.4. Pair Classification Tasks. Pair classification evaluates models on tasks such as entailment and contradiction. Our model achieves 71.03 after fine-tuning, up from 57.65 in the contrastive version. This strong improvement (+13.4 AP) underscores the effectiveness of task-specific adaptation in recognizing inter-sentence relational logic. Dataset-level scores for natural language inference tasks are summarized in Table A.4.

4.2.2.5. Semantic Textual Similarity (STS) Task. Our fine-tuned model scores 78.44 on STS, compared to 82.18 by `mpnet-base-v2`. STS requires high sensitivity to semantic nuance and often benefits from pretraining on NLI datasets. While our model performs well, it lacks the multilingual NLI exposure that gives multilingual models an edge in subtle textual alignment. STS performance on STS-B TR is reported in Table A.5 in the appendix.

4.2.2.6. Retrieval Tasks. Retrieval tasks show the widest score variance across the models. Our fine-tuned model achieves 52.75, behind the top models like `text-embedding-3-small` (64.99). Retrieval is asymmetric and benefits from instruction-tuned and retriever-optimized pretraining, which our model does not yet incorporate. We separate retrieval evaluations into two groups. Results for existing native and pre-translated Turkish datasets are reported in Table A.6, while performance on translated BEIR datasets is presented in Table A.7.

4.2.3. Impact of Contrastive Pretraining

Contrastive learning (via Cached-MNRL) substantially enhances model quality. Compared to the `bert-base-turkish-uncased` baseline, our model improves Mean(Task) by +10.3 points, showing that even weakly supervised data, if sufficiently filtered and semantically rich, enables robust embedding learning. Cached-MNRL simulates large batch sizes by caching embeddings across iterations, increasing the diversity

of in-batch negatives and improving generalization.

4.2.4. Effectiveness of Supervised Fine-Tuning

Supervised fine-tuning further boosts performance, especially on tasks requiring nuanced task-specific representations:

- Bitext Mining: Fine-tuning lifts performance from 37.7 to 86.88 (+49.2 F1).
- Pair Classification: Improves from 57.65 to 71.03 (+13.4 AP).
- STS: Rises from 74.93 to 78.44 (+3.5 Spearman).

These gains validate that while contrastive learning builds general-purpose embeddings, supervised tuning is essential to align them with downstream objectives.

4.2.5. Architectural and Capacity-Based Comparisons

The model architecture and scale remain the key determinants of success. Larger transformer encoders like `e5-large` and `gte-base` consistently outperform smaller ones. However, our 110M-parameter model closes much of the gap, suggesting that language-specific efficiency gains can partly offset scale disparities. In contrast, static word embeddings like `fasttext-turkish` and `glove-turkish` underperform in all tasks, affirming the superiority of contextual embeddings for modern NLP.

4.2.6. Cross-Lingual Alignment Limitations

Bitext mining results highlight a critical limitation: monolingual encoders struggle with cross-lingual tasks. Without multilingual pretraining or bilingual alignment objectives, embeddings in different languages occupy disjoint vector spaces, making cosine similarity ineffective.

To mitigate this, potential directions include:

- Pretraining on parallel corpora using bilingual contrastive objectives;
- Adding Translation Language Modeling (TLM) loss;
- Distillation from multilingual teacher models (e.g., LaBSE or mBERT).

These enhancements would help bridge the alignment gap in monolingual encoders and are necessary for strong performance in bitext tasks.

4.2.7. Analysis of Score Variability

As visualized in Figure 4.2, retrieval and bitext mining show the highest score spread across models, indicating sensitivity to training signals and multilingual context. Classification, clustering, and STS display more stable trends, reinforcing their status as easier-to-generalize tasks once core language representation is solid.

4.2.8. Summary of Findings

- Language-specific models can outperform multilingual ones in mono-task settings with sufficient training.
- Contrastive learning on weakly supervised corpora is highly effective if filtered properly.
- Supervised fine-tuning is critical for tasks with asymmetric or logic-based dependencies.
- Multilingual alignment remains a bottleneck for monolingual encoders in cross-lingual tasks.

These insights inform future development of sentence embeddings for Turkish and other morphologically rich, low-resource languages.

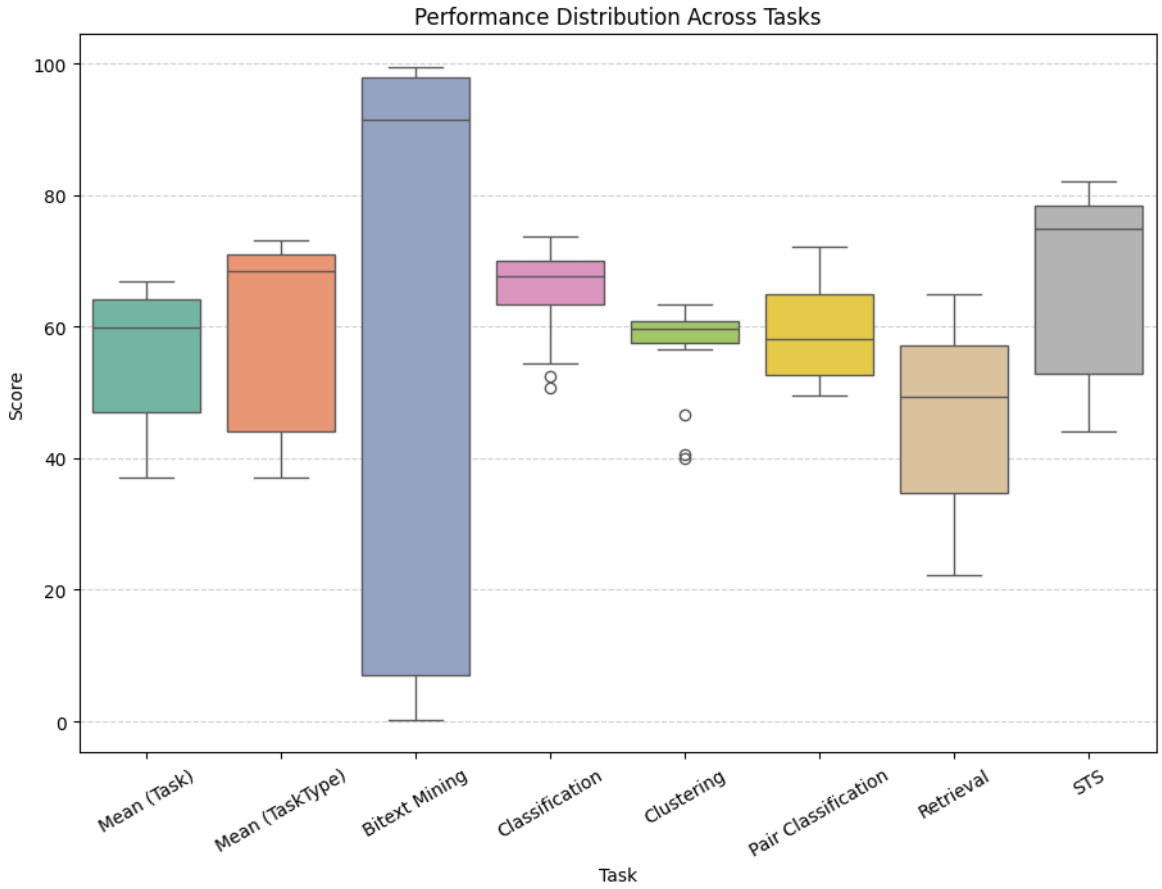


Figure 4.2. Distribution of model performance across six TR-MTEB task categories. Each box shows the score spread across models per task.

4.3. Limitations

Machine Translation Bias: A significant portion of TR-MTEB consists of translated datasets, especially for retrieval and pair classification. While these were validated via LLM-as-a-Judge, such datasets may still diverge from native Turkish in syntactic and semantic characteristics, potentially affecting generalizability. This approach mirrors similar practices in benchmarks like BEIR-PL and BEIR-NL, suggesting that such compromises are common but should be acknowledged.

Weak Supervision and Corpus Scale: Our 34.2M sentence pair corpus is sizable for Turkish but small compared to corpora used for English or Chinese models. Although filtered through embedding-based similarity thresholds, residual noise may

exist. Expanding the corpus and improving its coverage remains an open challenge.

Model Capacity Constraints: Our models are based on `bert-base-turkish-uncased` with 110M parameters. This restricts expressivity compared to models like `e5-large` (560M). Results show that careful pretraining and fine-tuning can mitigate this, but model size remains a ceiling on maximum achievable performance.

Task Coverage Gaps: While TR-MTEB spans six task types, it does not yet include summarization, re-ranking, or generation. Future extensions should incorporate these areas to reflect full real-world usage.

These limitations offer a roadmap for further improvements in both benchmark development and embedding model training for Turkish.

5. CONCLUSION

In this thesis, we presented TR-MTEB, the first large-scale, task-diverse benchmark designed specifically for evaluating Turkish sentence embedding models. Our benchmark includes 26 datasets spanning six core NLP task types: classification, clustering, pair classification, semantic textual similarity (STS), retrieval, and bitext mining. These datasets are drawn from both native Turkish sources and high-quality translated English resources, significantly extending the limited Turkish coverage found in existing multilingual evaluation suites like MTEB.

To facilitate the development of strong Turkish-specific embedding models, we constructed a 34.2 million–pair weakly supervised sentence corpus, curated from over 50 publicly available Turkish-language datasets. This corpus reflects domain diversity, including news, education, social media, and QA, and was carefully filtered using a combination of contrast trained and fine-tuned multilingual embedding models. The filtering pipeline was optimized using a combination of semantic similarity thresholds and manual evaluation to ensure quality while preserving coverage.

Using this corpus, we trained two Turkish sentence embedding models:

- A contrastively trained model based on `bert-base-turkish-uncased`, trained with Cached Multiple Negatives Ranking Loss (Cached-MNRL), achieving an effective batch size of 32,768 on a single A100 GPU setup.
- A supervised fine-tuned version of the above model, adapted using task-specific training data from TR-MTEB.

Our experimental results show that both models substantially outperform the baseline `bert-base-turkish-uncased` in all tasks. Particularly on monolingual tasks like classification and STS, our models perform competitively with multilingual SOTA models such as `multilingual-e5-base` and `gte-multilingual-base`, despite using

significantly fewer training resources (110M parameters vs. 305M–560M) and no English training data.

However, our results also reveal clear limitations. In cross-lingual tasks like bitext mining, the monolingual scope of our training data limits alignment capabilities with English, highlighting a persistent challenge for low-resource language models trained without multilingual signals. Furthermore, although our sentence pair corpus is the largest of its kind for Turkish, its scale is modest compared to corpora used in high-resource settings (e.g., 100M+ pairs used in BGE or E5). As such, there is still headroom to improve embedding quality through larger and more diverse data.

Additionally, our benchmark's reliance on machine-translated datasets introduces subtle risks in semantic fidelity, particularly for tasks involving nuance (e.g., STS or entailment). Despite manual curation and our LLM-as-a-Judge evaluation pipeline achieving 88.4% F1 alignment with human annotators, future iterations of TR-MTEB should incorporate more native datasets to minimize translation artifacts.

As future work, this study opens several promising directions:

- Scaling Data: Expanding the weakly supervised corpus with LLM-assisted filtering and multilingual alignment strategies.
- Multilingual Pretraining: Integrating Turkish–English training objectives to close the performance gap in bitext mining and zero-shot tasks.
- Generalization of Supervised Fine-Tuning: Investigating whether supervised fine-tuning on a subset of datasets within a task type can generalize and improve performance on other held-out datasets of the same task type. This would offer deeper insight into transferability and robustness of task-specific supervision.
- Architectural Improvements: Exploring larger model sizes or architectures (e.g., RoBERTa, ELECTRA) pretrained specifically for Turkish.
- Task Expansion: Extending TR-MTEB to include new task types such as summarization, reranking, and question generation to better reflect real-world appli-

cation scenarios.

- Robustness and Fairness: Evaluating how sentence embedding models perform across noisy inputs, different domains, and demographic subgroups in Turkish.

In summary, TR-MTEB provides a strong foundation for robust, reproducible, and fair evaluation of Turkish sentence embeddings. Our models and benchmark fill a critical gap for the Turkish NLP community, and we hope that the public release of all models, datasets, and evaluation pipelines will accelerate further progress in building inclusive NLP systems for low-resource and morphologically rich languages.

REFERENCES

1. Muennighoff, N., N. Tazi, L. Magne and N. Reimers, "MTEB: Massive Text Embedding Benchmark", A. Vlachos and I. Augenstein (Editors), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, Dubrovnik, Croatia, 2023.

2. Xiao, S., Z. Liu, P. Zhang, N. Muennighoff, D. Lian and J.-Y. Nie, "C-Pack: Packed Resources For General Chinese Embeddings", *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 641–649, New York, NY, USA, 2024.

3. Poswiata, R., S. Dadas and M. Perelkiewicz, "PL-MTEB: Polish Massive Text Embedding Benchmark", ArXiv:2405.10138, 2024.

4. Thakur, N., N. Reimers, A. Rücklé, A. Srivastava and I. Gurevych, "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models", *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, Virtual, 2021.

5. Wang, L., N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder and F. Wei, "Text Embeddings by Weakly-Supervised Contrastive Pre-training", ArXiv:2212.03533, 2022.

6. Li, Z., X. Zhang, Y. Zhang, D. Long, P. Xie and M. Zhang, "Towards General Text Embeddings with Multi-stage Contrastive Learning", ArXiv:2308.03281, 2023.

7. Schweter, S., 2020, "BERTurk - BERT Models for Turkish", `https://zenodo.org/records/3770924`, accessed: 6 July 2025.

8. Conneau, A. and D. Kiela, "SentEval: An Evaluation Toolkit for Universal Sentence Representations", N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis and T. Tokunaga (Editors), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan, 2018.

9. Cer, D., M. Diab, E. Agirre, I. Lopez-Gazpio and L. Specia, "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation", S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer and D. Jurgens (Editors), *Proceedings of the 11th International Workshop on Semantic Evaluation*, pp. 1–14, Vancouver, Canada, 2017.

10. Ciancone, M., I. Kerboua, M. Schaeffer and W. Siblini, "Extending the Massive Text Embedding Benchmark to French", ArXiv:2405.20468, 2024.

11. Snegirev, A., M. Tikhonova, A. Maksimova, A. Fenogenova and A. Abramov, "The Russian-focused embedders' exploration: ruMTEB benchmark and Russian embedding model design", ArXiv:2408.12503, 2024.

12. Enevoldsen, K., M. Kardos, N. Muennighoff and K. L. Nielbo, "The Scandinavian Embedding Benchmarks: Comprehensive Assessment of Multilingual and Monolingual Text Embedding", A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak and C. Zhang (Editors), *Advances in Neural Information Processing Systems*, Vol. 37, pp. 40336–40358, Vancouver, Canada, 2024.

13. Wojtasik, K., K. Wołowiec, V. Shishkin, A. Janz and M. Piasecki, "BEIR-PL: Zero Shot Information Retrieval Benchmark for the Polish Language", N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti and N. Xue (Editors), *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pp. 2149–2160, Torino, Italia, 2024.

14. Lotfi, E., N. Banar and W. Daelemans, "BEIR-NL: Zero-shot Information Retrieval Benchmark for the Dutch Language", S. Sharoff, A. R. Terryn, P. Zweigenbaum and R. Rapp (Editors), *Proceedings of the 18th Workshop on Building and Using Comparable Corpora (BUCC)*, pp. 36–45, Abu Dhabi, UAE, 2025.

15. Chen, J., S. Xiao, P. Zhang, K. Luo, D. Lian and Z. Liu, "M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation", L.-W. Ku, A. Martins and V. Srikumar (Editors), *Findings of the Association for Computational Linguistics*, pp. 2318–2335, Bangkok, Thailand, 2024.

16. Reimers, N. and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China, 2019.

17. Wang, L., N. Yang, X. Huang, L. Yang, R. Majumder and F. Wei, "Multilingual E5 Text Embeddings: A Technical Report", ArXiv:2402.05672, 2024.

18. Sezer, T., 2022, "THY-SA DataSet", `https://hackmd.io/@data-tdd/B1SyZQbyq`, accessed: 6 July 2025.

19. Sak, H., T. Güngör and M. Saraçlar, "Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus", B. Nordström and A. Ranta (Editors), *Advances in Natural Language Processing*, pp. 417–427, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

20. Karagoz, P., A. U. Ozturk and Y. Cemek, "IronyTR: Irony Detection in Turkish Informal Texts", *Int. J. Intell. Inf. Technol.*, Vol. 17, No. 4, p. 1–18, Oct. 2021.

21. Demirtas, E. and M. Pechenizkiy, "Cross-lingual Polarity Detection with Machine Translation", *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, New York, NY, USA, 2013.

22. Amasyalı, M. and T. Yıldırım, "Otomatik Haber Metinleri Sınıflandırma", pp. 224–226, Kuşadası, Turkey, 2004.

23. Çöltekin, C., "A Corpus of Turkish Offensive Language on Social Media", *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 6174–6184, Marseille, France, 2020.

24. Amasyalı, M. F. and B. Diri, "Automatic Turkish Text Categorization in Terms of Author, Genre and Gender", C. Kop, G. Fliedl, H. C. Mayr and E. Métais (Editors), *Natural Language Processing and Information Systems*, pp. 221–226, Springer Berlin Heidelberg, 2006.

25. Özturk, S., B. Sankur, T. Gungör, M. B. Yilmaz, B. Köroğlu, O. Ağin, M. İşbilen, Ulaş and M. Ahat, "Turkish Labeled Text Corpus", *22nd Signal Processing and*

Communications Applications Conference (SIU), pp. 1395–1398, Trabzon, Turkey, 2014.

26. Budur, E., R. Özçelik, T. Gungor and C. Potts, "Data and Representation for Turkish Natural Language Inference", B. Webber, T. Cohn, Y. He and Y. Liu (Editors), *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8253–8267, Online, 2020.

27. Conneau, A., R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk and V. Stoyanov, "XNLI: Evaluating Cross-lingual Sentence Representations", E. Riloff, D. Chiang, J. Hockenmaier and J. Tsujii (Editors), *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, Brussels, Belgium, 2018.

28. Beken Fikri, F., K. Oflazer and B. Yanikoglu, "Semantic Similarity Based Evaluation for Abstractive News Summarization", A. Bosselut, E. Durmus, V. P. Gangal, S. Gehrmann, Y. Jernite, L. Perez-Beltrachini, S. Shaikh and W. Xu (Editors), *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pp. 24–33, Online, 2021.

29. Budur, E., R. Özçelik, D. Soylu, O. Khattab, T. Güngör and C. Potts, "Building Efficient and Effective OpenQA Systems for Low-Resource Languages", *Knowledge-Based Systems*, Vol. 302, No. C, 2024.

30. Soygazi, F., O. Çiftçi, U. Kök and S. Cengiz, "THQuAD: Turkish Historic Question Answering Dataset for Reading Comprehension", *6th International Conference on Computer Science and Engineering (UBMK)*, pp. 215–220, Ankara, Turkey, 2021.

31. Artetxe, M., S. Ruder and D. Yogatama, "On the Cross-lingual Transferability of Monolingual Representations", ArXiv:1910.11856, 2019.

32. Kesgin, H. T., M. K. Yuce and M. F. Amasyali, "Developing and Evaluating Tiny to Medium-Sized Turkish BERT Models", ArXiv: 2307.14134, 2023.

33. Boteva, V., D. Gholipour, A. Sokolov and S. Riezler, "A Full-Text Learning to Rank Dataset for Medical Information Retrieval", N. Ferro, F. Crestani, M.-F.

Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff and G. Silvello (Editors), *Advances in Information Retrieval*, pp. 716–722, Padova, Italy, 2016.

34. Hoogeveen, D., K. M. Verspoor and T. Baldwin, "CQADupStack: A Benchmark Data Set for Community Question-Answering Research", *Proceedings of the 20th Australasian Document Computing Symposium (ADCS)*, ADCS '15, pp. 3:1–3:8, New York, NY, USA, 2015.

35. Kaggle, 2017, "Quora Question Pairs", `https://kaggle.com/competitions/quora-question-pairs`, accessed: 6 July 2025.

36. Cohan, A., S. Feldman, I. Beltagy, D. Downey and D. Weld, "SPECTER: Document-level Representation Learning using Citation-informed Transformers", D. Jurafsky, J. Chai, N. Schluter and J. Tetreault (Editors), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2270–2282, Online, 2020.

37. Wadden, D., K. Lo, B. Kuehl, A. Cohan, I. Beltagy, L. L. Wang and H. Hajishirzi, "SciFact-Open: Towards Open-domain Scientific Claim Verification", Y. Goldberg, Z. Kozareva and Y. Zhang (Editors), *Findings of the Association for Computational Linguistics: EMNLP*, pp. 4719–4734, Abu Dhabi, United Arab Emirates, 2022.

38. Dang, J., S. Singh, D. D'souza, A. Ahmadian, A. Salamanca, M. Smith, A. Peppin, S. Hong, M. Govindassamy, T. Zhao, S. Kublik, M. Amer, V. Aryabumi, J. A. Campos, Y.-C. Tan, T. Kocmi, F. Strub, N. Grinsztajn, Y. Flet-Berliac, A. Locatelli, H. Lin, D. Talupuru, B. Venkitesh, D. Cairuz, B. Yang, T. Chung, W.-Y. Ko, S. S. Shi, A. Shukayev, S. Bae, A. Piktus, R. Castagné, F. Cruz-Salinas, E. Kim, L. Crawhall-Stein, A. Morisot, S. Roy, P. Blunsom, I. Zhang, A. Gomez, N. Frosst, M. Fadaee, B. Ermis, A. Üstün and S. Hooker, "Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier", ArXiv:2412.04261, 2024.

39. Wikimedia Foundation, 2024, "Wikimedia Downloads", `https://dumps.wikimedia.org`, accessed: 6 July 2025.

40. De Bruyn, M., E. Lotfi, J. Buhmann and W. Daelemans, "MFAQ: a Multilingual FAQ Dataset", *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pp. 1–13, Punta Cana, Dominican Republic, 2021.

41. Scherrer, Y., 2020, "TaPaCo: A Corpus of Sentential Paraphrases for 73 Languages", `https://zenodo.org/record/3707949`, accessed: 6 July 2025.

42. Ladhak, F., E. Durmus, C. Cardie and K. McKeown, "WikiLingua: A New Benchmark Dataset for Cross-Lingual Abstractive Summarization", T. Cohn, Y. He and Y. Liu (Editors), *Findings of the Association for Computational Linguistics: EMNLP*, pp. 4034–4048, Online, 2020.

43. Gultekin, M., 2023, "Wikipedia Turkish Summarization Dataset", `https://huggingface.co/datasets/musabg/wikipedia-tr-summarization`, accessed: 6 July 2025.

44. Vasilyev, O., F. Isono and J. Bohannon, "Linear Cross-Lingual Mapping of Sentence Embeddings", L.-W. Ku, A. Martins and V. Srikumar (Editors), *Findings of the Association for Computational Linguistics*, pp. 8163–8171, Bangkok, Thailand, 2024.

45. Lian, W., B. Goodson, E. Pentland, A. Cook, C. Vong and Teknium, 2023, "OpenOrca: An Open Dataset of GPT Augmented FLAN Reasoning Traces", `https://huggingface.co/datasets/Open-Orca/OpenOrca`, accessed: 6 July 2025.

46. Clark, P., I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick and O. Tafjord, "Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge", ArXiv:1803.05457, 2018.

47. Leeb, F. and B. Schölkopf, "A Diverse Multilingual News Headlines Dataset From Around the World", ArXiv:2403.19352, 2024.

48. Unal, M. E., B. Citamak, S. Yagcioglu, A. Erdem, E. Erdem, N. I. Cinbis and R. Cakici, "TasvirEt: A Benchmark Dataset for Automatic Turkish Description

Generation from Images", *24th Signal Processing and Communication Application Conference (SIU)*, pp. 1977–1980, Zonguldak, Turkey, 2016.

49. Bayram, M., 2024, "Türkçe Tıbbi Soru-Cevap Veri Seti: 167 Bin Sağlık Sorusu ve Cevabı", `https://zenodo.org/record/12770916`, accessed: 6 July 2025.

50. Bayram, A., 2024, "Onedio Haberler Veri Seti", `https://huggingface.co/datasets/alibayram/onedio_haberler`, accessed: 6 July 2025.

51. Üyük, C., D. Rovó, S. Shaghayeghkolli, R. Varol, G. Groh and D. Dementieva, "Crafting Tomorrow's Headlines: Neural News Generation and Detection in English, Turkish, Hungarian, and Persian", D. Dementieva, O. Ignat, Z. Jin, R. Mihalcea, G. Piatti, J. Tetreault, S. Wilson and J. Zhao (Editors), *Proceedings of the Third Workshop on NLP for Positive Impact*, pp. 271–307, Miami, Florida, USA, 2024.

52. Bezir, A., 2024, "gsm8k-tr", `https://huggingface.co/datasets/bezir/gsm8k-tr`, accessed: 6 July 2025.

53. Altinok, D., 2024, "InstrucTurca: A Diverse Instructional Content Dataset for Turkish", `https://huggingface.co/datasets/turkish-nlp-suite/InstrucTurca`, accessed: 6 July 2025.

54. Gureja, S., L. Miranda, S. Islam, R. Maheshwary, D. Sharma, G. Winata, N. Lambert, S. Ruder, S. Hooker and M. Fadaee, "M-RewardBench: Evaluating Reward Models in Multilingual Settings", ArXiv:2410.15522, 2024.

55. Bayram, M., 2024, "Turkish MMLU: Yapay Zeka ve Akademik Uygulamalar İçin En Kapsamlı ve Özgün Türkçe Veri Seti", `https://zenodo.org/record/13378019`, accessed: 6 July 2025.

56. İncidelen, M. and M. Aydoğan, "Developing Question-Answering Models in Low-Resource Languages: A Case Study on Turkish Medical Texts Using Transformer-Based Approaches", *8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, pp. 1–4, Malatya, Turkey, 2024.

57. Clark, E., S. Rijhwani, S. Gehrmann, J. Maynez, R. Aharoni, V. Nikolaev, T. Sellam, A. Siddhant, D. Das and A. Parikh, "SEAHORSE: A Multilingual, Multifaceted Dataset for Summarization Evaluation", ArXiv:22305.13194, 2023.

58. Palen-Michel, C. and C. Lignos, "LR-Sum: Summarization for Less-Resourced Languages", A. Rogers, J. Boyd-Graber and N. Okazaki (Editors), *Findings of the Association for Computational Linguistics: ACL*, pp. 6829–6844, Toronto, Canada, 2023.

59. Iana, A., G. Glavaš and H. Paulheim, "Mind Your Language: A Multilingual Dataset for Cross-Lingual News Recommendation", *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 553–563, Washington DC, USA, 2024.

60. Bulut, M., 2024, "Patient Doctor Q&A TR", `https://huggingface.co/datasets/kayrab/patient-doctor-qa-tr-19583`, accessed: 6 July 2025.

61. sekerlipencere, 2024, "zynpdata: Türkiye'nin En Büyük Açık Kaynaklı Türkçe Forum Veri Seti", `https://github.com/sekerlipencere/zynpdata-zynp_ai-teknofest`, accessed: 6 July 2025.

62. Scialom, T., P.-A. Dray, S. Lamprier, B. Piwowarski and J. Staiano, "MLSUM: The Multilingual Summarization Corpus", B. Webber, T. Cohn, Y. He and Y. Liu (Editors), *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8051–8067, Online, 2020.

63. Baykara, B. and T. Güngör, "Abstractive text summarization and new large-scale datasets for agglutinative languages Turkish and Hungarian", *Lang. Resour. Eval.*, Vol. 56, No. 3, p. 973–1007, 2022.

64. Özturk, S., B. Sankur, T. Gungör, M. B. Yilmaz, B. Köroğlu, O. Ağin, M. İşbilen, Ç. Ulaş and M. Ahat, "Turkish Labeled Text Corpus", *22nd Signal Processing and Communications Applications Conference (SIU)*, pp. 1395–1398, Trabzon, Turkey, 2014.

65. Gao, L., Y. Zhang, J. Han and J. Callan, "Scaling Deep Contrastive Learning Batch Size Under Memory Limited Setup", ArXiv:2101.06983, 2021.

66. Su, J., 2022, "CoSENT: A More Efficient Sentence Vector Scheme Than Sentence-BERT", `https://kexue.fm/archives/8847`, accessed: 6 July 2025.

67. Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest and A. Rush, "Transformers: State-of-the-Art Natural Language Processing", Q. Liu and D. Schlangen (Editors), *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, 2020.

68. Lhoest, Q., A. Villanova del Moral, Y. Jernite, A. Thakur, P. von Platen, S. Patil, J. Chaumond, M. Drame, J. Plu, L. Tunstall, J. Davison, M. Šaško, G. Chhablani, B. Malik, S. Brandeis, T. Le Scao, V. Sanh, C. Xu, N. Patry, A. McMillan-Major, P. Schmid, S. Gugger, C. Delangue, T. Matussière, L. Debut, S. Bekman, P. Cistac, T. Goehringer, V. Mustar, F. Lagunas, A. Rush and T. Wolf, "Datasets: A Community Library for Natural Language Processing", H. Adel and S. Shi (Editors), *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 175–184, Online and Punta Cana, Dominican Republic, 2021.

69. Feng, F., Y. Yang, D. Cer, N. Arivazhagan and W. Wang, "Language-agnostic BERT Sentence Embedding", S. Muresan, P. Nakov and A. Villavicencio (Editors), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 878–891, Dublin, Ireland, 2022.

70. Bojanowski, P., E. Grave, A. Joulin and T. Mikolov, "Enriching Word Vectors With Subword Information", ArXiv:1607.04606, 2016.

71. Sarıtaş, K., C. A. Öz and T. Güngör, "A Comprehensive Analysis of Static Word Embeddings for Turkish", *Expert Systems with Applications*, Vol. 252, pp. 123–124, 2024.

APPENDIX A: MTEB-TR Task Results

A.1. Classification Results

Table A.1 reports classification results across all evaluated models in TR-MTEB. This task type assesses the ability of sentence embeddings to support downstream classification objectives via logistic regression. Datasets span diverse domains, including sentiment, news, and offensive language detection.

A.2. Clustering Results

Table A.2 presents the performance of embedding models on two clustering datasets in Turkish. These datasets require grouping based on authorship or academic discipline, and models are evaluated via V-Measure.

A.3. Bitext Mining Results

Table A.3 shows the performance of models on the WMT16 EN–TR bitext mining task. This task measures the model's cross-lingual alignment capability. As expected, multilingual models outperform monolingual ones.

A.4. Pair Classification Results

As shown in Table A.4, we include three natural language inference datasets: XNLI, SNLI-TR, and MultiNLI-TR. These tasks evaluate whether embeddings can distinguish semantic entailment relations between sentence pairs.

A.5. Semantic Textual Similarity Results

Table A.5 summarizes results for the STS-B TR dataset. Models are evaluated by computing cosine similarity between sentence embeddings and correlating with human-annotated similarity scores using Spearman and Pearson metrics.

A.6. Retrieval Results

Table A.6 and Table A.7 present retrieval task results across all models in TR-MTEB. The first table includes datasets that existed in Turkish or were previously translated (e.g., MS MARCO TR, SQuAD TR), while the second table reports performance on seven datasets newly translated into Turkish as part of this work. These tasks evaluate models on dense passage retrieval settings, with nDCG@10 used as the main metric.

Table A.1. Classification task performance across models in TR-MTEB. Scores reflect macro-average F1 across categories.

Model	THY Sentiment Classification	TS Timeline News Category Classification	Turkish 75 News Classification	Turkish Movie Sentiment Classification	Turkish News Category Classification	Turkish Offensive Language Classification	Turkish Product Sentiment Classification
multilingual-e5-large	64.17	64.87	93.33	80.64	92.00	61.40	62.29
multilingual-e5-large-instruct	66.15	65.02	92.67	85.25	93.76	60.69	67.23
gte-multilingual-base	59.37	64.26	84.00	70.78	90.20	58.70	62.06
turkish-embedding-model-fine-tuned	62.31	60.67	95.33	78.19	91.92	64.62	61.12
text-embedding-3-small	59.07	64.14	98.67	71.86	92.16	60.78	56.79
multilingual-e5-base	62.63	62.65	84.00	72.60	90.80	57.81	60.61
turkish-embedding-model	63.81	66.60	89.33	78.91	85.92	65.23	58.85
paraphrase-multilingual-mpnet-base-v2	59.55	59.34	94.00	75.94	86.48	60.46	62.29
multilingual-e5-small	55.88	60.30	82.00	67.77	90.36	59.13	58.03
paraphrase-multilingual-MiniLM-L12-v2	57.87	53.20	88.00	70.02	85.88	58.08	59.28
LaBSE	59.61	48.81	89.33	66.94	86.16	59.26	59.29
bert-base-turkish-uncased	66.01	62.85	88.67	71.34	90.08	64.35	61.99
fasttext	56.11	46.16	86.67	64.92	86.32	58.20	55.83
glove	54.58	51.07	78.67	63.65	87.32	57.41	55.15
all-MiniLM-L12-v2	50.72	23.85	58.67	56.56	57.48	52.99	52.89
all-MiniLM-L6-v2	49.98	23.29	71.33	57.23	61.52	51.33	52.23
all-mpnet-base-v2	47.81	29.84	78.67	54.73	68.36	52.66	52.00

Table A.2. Performance of embedding models on Turkish clustering tasks. Evaluation based on V-measure for two domains: academic abstracts and opinion columns.

Model	**Turkish Abstract Corpus Clustering**	**Turkish Column Writing Clustering**
multilingual-e5-large	59.88	60.96
multilingual-e5-large-instruct	62.07	61.50
gte-multilingual-base	59.01	61.78
turkish-embedding-model-fine-tuned	59.73	63.26
text-embedding-3-small	60.63	63.52
multilingual-e5-base	58.53	61.96
turkish-embedding-model	63.58	63.31
paraphrase-multilingual-mpnet-base-v2	57.62	57.43
multilingual-e5-small	56.64	61.63
paraphrase-multilingual-MiniLM-L12-v2	56.17	57.11
LaBSE	55.37	61.44
bert-base-turkish-uncased	54.55	66.98
fasttext	50.91	68.58
glove	51.84	64.99
all-MiniLM-L12-v2	30.61	49.28
all-MiniLM-L6-v2	31.01	50.13
all-mpnet-base-v2	39.80	53.46

Table A.3. Performance of embedding models on Turkish–English bitext mining using the WMT16 EN-TR dataset. Evaluation based on F1-score for top-1 retrieval.

Model	WMT16 EN-TR
multilingual-e5-large	99.43
multilingual-e5-large-instruct	98.99
gte-multilingual-base	97.98
turkish-embedding-model-fine-tuned	86.88
text-embedding-3-small	91.55
multilingual-e5-base	98.51
turkish-embedding-model	37.70
paraphrase-multilingual-mpnet-base-v2	96.06
multilingual-e5-small	97.46
paraphrase-multilingual-MiniLM-L12-v2	93.67
LaBSE	99.53
bert-base-turkish-uncased	7.07
fasttext	4.52
glove	0.22
all-MiniLM-L12-v2	8.27
all-MiniLM-L6-v2	6.78
all-mpnet-base-v2	5.07

Table A.4. Performance of embedding models on pair classification tasks in Turkish using XNLI, SNLI TR, and MultiNLI-TR datasets. Evaluation uses macro-average F1 over entailment detection.

Model	XNLI	SNLI-TR	MultiNLI-TR
multilingual-e5-large	73.91	55.84	65.28
multilingual-e5-large-instruct	73.45	52.47	65.28
gte-multilingual-base	76.59	60.26	65.38
turkish-embedding-model-fine-tuned	79.60	63.18	70.31
text-embedding-3-small	66.18	50.49	57.71
multilingual-e5-base	70.15	53.98	62.47
turkish-embedding-model	64.30	51.35	57.29
paraphrase-multilingual-mpnet-base-v2	84.25	60.67	71.74
multilingual-e5-small	70.20	55.06	61.74
paraphrase-multilingual-MiniLM-L12-v2	79.84	59.72	69.13
LaBSE	65.63	47.61	60.25
bert-base-turkish-uncased	61.79	42.98	53.98
fasttext	59.82	46.04	52.23
glove	55.79	45.74	46.90
all-MiniLM-L12-v2	58.48	45.61	51.27
all-MiniLM-L6-v2	57.47	44.19	49.91
all-mpnet-base-v2	56.06	43.74	49.58

Table A.5. Semantic similarity performance of embedding models on the Turkish version of the STS-Benchmark dataset. Evaluation based on Spearman correlation.

Model	STS-B TR
multilingual-e5-large	81.18
multilingual-e5-large-instruct	81.23
gte-multilingual-base	80.68
turkish-embedding-model-fine-tuned	78.44
text-embedding-3-small	70.85
multilingual-e5-base	78.37
turkish-embedding-model	74.93
paraphrase-multilingual-mpnet-base-v2	82.18
multilingual-e5-small	76.18
paraphrase-multilingual-MiniLM-L12-v2	78.41
LaBSE	71.75
bert-base-turkish-uncased	56.74
fasttext	50.08
glove	44.12
all-MiniLM-L12-v2	52.78
all-MiniLM-L6-v2	48.55
all-mpnet-base-v2	50.21

Table A.6. Retrieval task results on Turkish or previously translated datasets such as MS MARCO TR, SQuAD TR, TQuAD, and XQuAD. Evaluation is based on nDCG@10.

Model	MS Marco TR	SQuAD TR Retrieval	TQuAD Retrieval	Turkish Product Sentiment Classification
multilingual-e5-large	58.10	80.13	88.61	96.73
multilingual-e5-large-instruct	56.78	73.01	84.68	94.66
gte-multilingual-base	56.84	68.64	81.91	92.29
turkish-embedding-model-fine-tuned	57.19	58.14	80.41	86.99
text-embedding-3-small	53.37	65.56	79.53	90.34
multilingual-e5-base	57.14	76.90	86.92	95.11
turkish-embedding-model	54.35	54.11	74.83	85.44
paraphrase-multilingual-mpnet-base-v2	51.54	58.03	65.89	85.95
multilingual-e5-small	56.46	73.93	88.83	94.33
paraphrase-multilingual-MiniLM-L12-v2	47.06	55.11	64.01	82.83
LaBSE	46.93	58.36	66.25	83.17
bert-base-turkish-uncased	34.68	44.49	60.56	72.37
fasttext	31.59	44.39	65.55	71.80
glove	20.47	27.04	42.55	50.60
all-MiniLM-L12-v2	32.03	20.55	43.55	49.58
all-MiniLM-L6-v2	27.32	14.47	36.35	40.03
all-mpnet-base-v2	25.46	12.81	27.46	36.53

Table A.7. Retrieval task results on Turkish versions of BEIR datasets translated in this work. Includes ArguAna TR, FiQA TR, CQADupstack Gaming TR, NFCorpus TR, and others. Evaluation metric is nDCG@10.

Model	ArguAna TR	CQADupstack Gaming TR	FiQA2018 TR	NFCorpus TR	Quora TR	SCIDOCS TR	SciFact TR
multilingual-e5-large	47.34	61.12	44.60	8.02	95.88	3.14	83.10
multilingual-e5-large-instruct	44.90	51.26	36.94	8.35	95.89	2.81	80.00
gte-multilingual-base	47.72	56.26	40.57	7.79	95.13	3.93	81.58
turkish-embedding-model-fine-tuned	36.11	51.72	41.65	8.68	94.16	2.22	62.94
text-embedding-3-small	40.56	80.73	73.11	22.90	98.85	16.27	93.64
multilingual-e5-base	44.80	56.05	37.94	8.33	95.23	3.07	79.75
turkish-embedding-model	38.89	54.03	39.02	9.36	94.65	2.88	70.33
paraphrase-multilingual-mpnet-base-v2	39.34	46.16	27.92	5.50	94.70	3.18	63.80
multilingual-e5-small	42.73	53.29	32.57	7.90	94.19	2.09	75.52
paraphrase-multilingual-MiniLM-L12-v2	31.97	42.38	25.45	5.83	93.68	2.88	61.11
LaBSE	33.49	39.72	21.94	6.08	92.75	2.64	59.81
bert-base-turkish-uncased	20.62	23.36	11.21	2.58	86.23	1.04	25.87
fasttext	11.57	23.73	11.55	2.86	75.66	0.63	42.00
glove	8.38	16.43	9.55	1.14	56.74	0.91	31.79
all-MiniLM-L12-v2	1.25	28.84	11.95	2.71	79.03	1.54	32.59
all-MiniLM-L6-v2	1.38	22.24	4.98	3.88	73.07	0.94	31.15
all-mpnet-base-v2	2.35	22.63	7.14	2.63	74.77	1.37	31.69