Tooka-SBERT: Sentence Embedding model for Persian

Anonymous ACL submission

Abstract

We introduce Tooka-SBERT, a family of Persian sentence embedding models designed to enhance semantic understanding for Persian. The models are released in two sizes—Small (123M parameters) and Large (353M parameters)—both built upon the TookaBERT backbone. Tooka-SBERT is pretrained on the Targoman News corpus and fine-tuned using high-quality synthetic Persian sentence pair datasets to improve semantic alignment. We evaluate Tooka-SBERT on PTEB, a Persian adaptation of the MTEB benchmark, where the Large model achieves an average score of 70.54% and the Small model 69.49%, outperforming some strong multilingual baselines. Tooka-SBERT provides a compact and high-performing open-source solution for Persian sentence representation, with efficient inference suitable for both GPU and CPU environments.

1 Introduction

Text embeddings are a foundational component in natural language processing, powering a wide array of applications such as clustering, search systems, text mining, and serving as feature representations for downstream models (Wang et al., 2024). Their ability to convert semantic relationships into spatial relationships between vectors is crucial for efficient information retrieval systems and language models.

With the rapid adoption of Large Language Models and growing concerns about hallucinations, Retrieval Augmented Generation (RAG) has emerged as a critical approach to enhance factual accuracy (Lewis et al., 2020). RAG pipelines rely heavily on robust embedding models that can accurately capture semantic similarity and retrieve the most relevant information. These models must not only offer strong semantic alignment, but also be computationally efficient to enable fast inference and retrieval in real-world pipelines. Sentence-BERT (Reimers and Gurevych, 2019) introduced a paradigm for generating independent, high-quality sentence embeddings, making it particularly effective for retrieval tasks. However, for Persian language applications, the scarcity of robust embedding models poses a challenge, making the development of high-performing Persian embeddings essential for advancing Persian RAG systems.

This paper introduces Tooka-SBERT, a family of text embedding models, designed specifically for semantic textual similarity and embedding tasks in Persian. These models map sentences and paragraphs to a dense vector space where semantically similar texts are positioned closely together. The Tooka-SBERT-V2 model is available in two sizes: Small (123M parameters) and Large (353M parameters). Our models are built upon TookaBERT (SadraeiJavaheri et al., 2024), a Persian pre-trained language model.

Our main contributions are as follows:

- We introduce Tooka-SBERT, a family of compact sentence embedding models for Persian. Despite having relatively few parameters, Tooka-SBERT models achieve strong performance across diverse tasks in Persian. The Large-V2 variant outperforms state-of-the-art baselines, achieving approximately 1.2% higher performance than multilingual-e5-base (Wang et al., 2024) and around 3.5% higher than Qwen3-Embedding-0.6B (Zhang et al., 2025). The Small-V2 variant, with fewer parameters, also surpasses multilingual-e5-base on the

PTEB benchmark.

- We present the PTEB benchmark, a Persian adaptation of MTEB (Muennighoff et al., 2022), constructed by collecting and curating datasets across a range of tasks to enable comprehensive evaluation of Persian sentence embeddings.

2 Related Works

2.1 General Text Embeddings

The study of text embeddings has evolved significantly, beginning with statistical and matrix-based techniques before advancing to neural and transformer-based architectures. Early approaches such as Latent Semantic Indexing (LSI) (Deerwester et al., 1990) and Latent Dirichlet Allocation (Blei et al., 2001) represented documents via word co-occurrence or topic distributions. Later, neural methods using word embeddings such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) introduced more efficient vector representations for semantic similarity, but they lacked context awareness. The field advanced significantly with deep learning-based contextualized models, such as ELMo (Peters et al., 2018) and transformer-based architectures like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019).

While BERT and RoBERTa set new state-of-the-art performance on sentence-pair regression tasks like semantic textual similarity (STS), they require feeding both sentences into the network, leading to significant computational overhead. To overcome this, Sentence-BERT (Reimers and Gurevych, 2019) was introduced, which uses siamese and triplet network structures to derive fixed-size, semantically meaningful sentence embeddings that can be compared efficiently using cosine similarity. Contrastive learning methods, such as SimCSE (Gao et al., 2021b), have further advanced general-purpose text representations by fine-tuning transformers on positive and negative text pairs using a contrastive loss objective. However, models like SimCSE were primarily trained on single tasks and were not inherently suitable for broader applications. This led to the development of a new generation of models designed to generalize across a wider range of tasks, including retrieval, classification, and question-answering. Training these models often involves multi-stage and multi-task fine-tuning strategies that incorporate weakly-supervised contrastive training. Techniques like AliBi (Press et al., 2022) and Rotary Position Embeddings (RoPE) (Su et al., 2024) have enabled models to handle longer text sequences, while Matryoshka Representation Learning (Kusupati et al., 2022) allows for truncating embeddings to smaller dimensions without significantly compromising performance.

2.2 Multilingual Embedding Models

The development of multilingual models has been crucial for extending NLP capabilities beyond English. Early examples include Multilingual BERT (mBERT) (Devlin et al., 2019) and XLM-RoBERTa (XLM-R) (Conneau et al., 2020), which were trained on large corpora spanning many languages. Also, multilingual embedding models have advanced the field through novel architectures and training strategies. Multilingual E5 (mE5) extends the English E5 framework with a two-stage approach: weakly-supervised contrastive pre-training on billions of multilingual text pairs, followed by supervised fine-tuning on labeled datasets. BGE M3 (Chen et al., 2024), built on XLM-R, supports long input sequences and utilizes RetroMAE pretraining (Xiao et al., 2022) along with a multi-CLS pooling mechanism. It undergoes contrastive training on unlabeled pairs, followed by fine-tuning on task-specific labeled data. Similarly, Jina-embeddings-v3 (Sturua et al., 2025), also based on XLM-RoBERTa, leverages RoPE positional encoding and LoRA adapters (Hu et al., 2022) for long-context multilingual retrieval, achieving strong performance on MTEB tasks. The Qwen3 (Zhang et al., 2025) Embedding series employs a multi-stage training pipeline with LLM-generated synthetic data, robust model merging strategies, and fine-tuning.

2.3 Persian Embedding Models

Persian remains significantly underrepresented in large-scale text embedding research. While several open-source models have been released on HuggingFace by the Persian NLP

Dataset	#Train	#Test	Structure
News	16M	1.8M	(title, subtitle, text)
NLI	68K	7.5K	(sentence, paraphrases, entailment, neutral, contradiction)
RAFT	103K	11K	(question, oracle_context, answer, negative_contexts)
MIRACL	20K	2.2K	(query, doc, score, relevance)

Table 1: Overview of datasets used for training and evaluation.

community, they generally lag behind models developed for high-resource languages in terms of performance. One of the early efforts to adapt Sentence-Transformer architectures for Persian was PersianSentenceTransformers (Farahani, 2020), which leveraged ParsBERT (Farahani et al., 2020) and was fine-tuned on FarsTail (Amirkhani et al., 2023)— the first Persian NLI dataset—as well as a modified Wikipedia-Triplet-Sections approach (Ein Dor et al., 2018), which involved extensive preprocessing and filtering of Wikipedia articles to generate meaningful triplets and Similar/Dissimilar sentence pairs. Another ParsBERT-based model, sentence-transformer-parsbert-fa (Ahmadi, 2024), was trained specifically to enhance Retrieval-Augmented Generation systems for applications such as QA and chatbots.

Building on cross-lingual architectures, Sobhi (2024) fine-tuned XLM-RoBERTa-Large on a variety of Persian datasets, including ParsiNLU (Khashabi et al., 2021) and PQuAD (Darvishi et al., 2023), to support different tasks. Similarly, Heydari (2024) fine-tuned XLM-RoBERTa-Base on a large-scale Persian corpus to produce high-quality contextual embeddings for both monolingual and multilingual applications. The maux-gte-persian model (Mirzaei, 2024), derived from GTE-multilingual (Zhang et al., 2024), was fine-tuned using Persian sentence pairs translated from English with GPT-4, offering strong performance across Persian semantic tasks. Finally, Hakim (Sarmadi et al., 2025) stands out as a purpose-built Persian embedding model that applies the RetroMAE architecture in a two-stage contrastive and supervised training pipeline. It applies task-specific instructions and dedicated CLS-token supervision.

3 Training Data

To ensure strong performance across various tasks, we utilized a combination of existing and synthetic datasets. *Targoman Large Persian Corpus* (TLPC) (Targoman, 2022) is the largest among them. It was collected by scraping over 800 popular Persian websites, resulting in more than 75 million documents across diverse domains. We used its News section and, after normalization, extracted the title, subtitle, and main text as training data. TLPC is released under the CC-BY-NC-SA-4.0 license, and we used it strictly for non-commercial model training.

NLI is a synthetic dataset generated by an LLM model. For each input sentence, the model generated a tuple containing paraphrases, as well as entailment, neutral, and contradiction sentences.

Another synthetic dataset, *RAFT*, was generated by LLMs using webpages crawled from Wikipedia. Each webpage was split into multiple chunks; one chunk was selected as the oracle context, and the LLM was prompted to generate a corresponding question and answer. The remaining chunks were treated as negative contexts.

MIRACL (Zhang et al., 2023) is a multilingual retrieval dataset in which each sample consists of a query, a document, and a binary relevance label (1 for relevant, 0 for irrelevant). We used its training split to fine-tune our model. The dataset is released under the Apache License 2.0, which permits both commercial and non-commercial use with attribution. Additionally, we used a cross-encoder model, `bge-reranker-v2-m3` (Chen et al., 2024), to compute a continuous similarity score between the query and document pairs, which was used as a soft supervision signal during training.

Table 1 summarizes the datasets used during training.

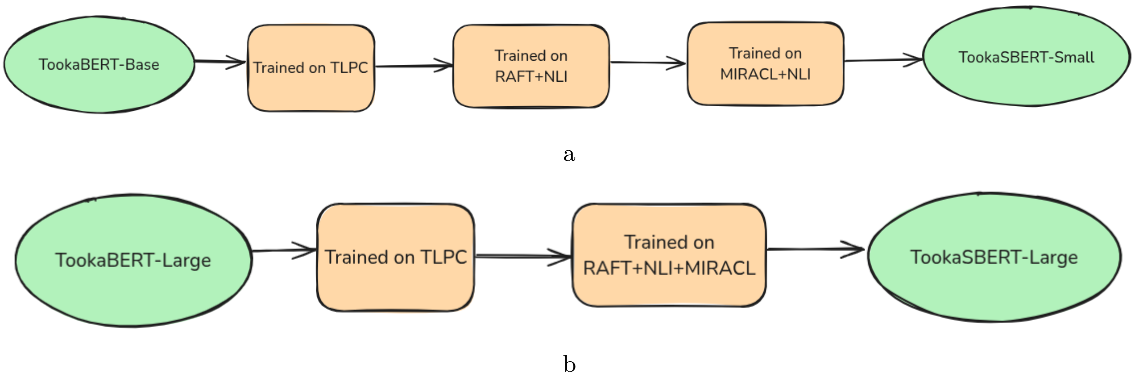

Figure 1: Training pipelines of Tooka-SBERT models. (a) Tooka-SBERT-Small, and (b) Tooka-SBERT-Large.

4 Methodology

4.1 Overview

We conducted a series of experiments to develop high-quality sentence embeddings for the Persian language. Our goal was to train a model that performs well across a variety of semantic tasks such as semantic textual similarity (STS), information retrieval, reranking, and classification.

Through these experiments, we explored different training strategies. We released the first successful result as Tooka-SBERT-V1. However, our main contribution in this work is Tooka-SBERT-V2, a more robust and versatile model trained using multi-stage techniques. We trained our model in two sizes: Small (123M parameters) and Large (353M parameters).

The training strategy for V2 consists of two main stages:

1. Warming-up on TLPC data
2. Fine-tuning on a collection of datasets

We implemented our training pipeline using the `sentence-transformers` library, which provides flexible support for various loss functions, training strategies, and efficient multi-dataset handling.

4.2 Tooka-SBERT-Small

We use TookaBERT-Base as the backbone for our Small model. The training process follows a multi-stage strategy, illustrated in Fig. 1a

Stage 1 – Warm-up on TLPC: We pretrain the model on the Persian news dataset provided by Targoman, using an asymmetric input format to differentiate between query and document pairs. Specifically, we prepend:

- «سوال:» ("question:") to query sentences
- «متن:» ("text:") to document passages

This prefix-based formatting helps the model distinguish between queries and documents, improving its ability to capture semantic relationships between questions and relevant texts. (Wang et al., 2022)

We primarily employed *Cached Multiple Negatives Ranking Loss* (Gao et al., 2021a) during training. This loss function is widely used in sentence embedding models, particularly for contrastive learning in retrieval settings. It maximizes the similarity of a query and its corresponding positive while minimizing the similarity with all in-batch negatives. Unlike traditional triplet losses, it doesn't require explicit hard negative mining, making it more efficient and stable in large-scale training. Furthermore, *Cached Multiple Negatives Ranking Loss* allows training with effectively larger batch sizes without the need for additional VRAM, whereas *Contrastive Loss* (Radford et al., 2021) typically requires very large batch sizes to achieve good convergence.

Stage 2 - Fine-Tuning on RAFT + NLI: We used a proportional sampling strategy across the Raft and NLI datasets, training for 5 epochs with *Cached Multiple Negatives Ranking Loss*. Sampling proportion was based on dataset size to ensure balanced coverage.

- RAFT Format: (question, oracle_context,

Dataset	Structure	Loss Function
MIRACL	(query, doc, score — float)	CoSENTLoss
	(query, doc, relevance — binary)	OnlineContrastiveLoss
NLI	(sentence, contradiction, 0)	SoftmaxLoss
	(sentence, neutral, 1)	
	(sentence, entailment, 2)	
	(sentence, paraphrase, contradiction)	CMNRLoss
RAFT	(question, oracle, negative_1, negative_2, negative_3)	CMNRLoss
	(question, answer)	CMNRLoss

Table 2: Input formats and corresponding loss functions used for each dataset during training the Large model. **CMNRLoss** is *Cached Multiple Negatives Ranking Loss*.

negative_context$_1$, negative_context$_2$, negative_context$_3$)

- NLI Format: sentence, paraphrase, contradiction

Stage 3 – Fine-tuning on MIRACL + NLI: We applied a round-robin strategy for 170 steps, where batches are sampled alternately from each dataset until one is exhausted. This means not all samples from each dataset may be used, but sampling is performed equally across datasets. The MIRACL dataset was trained using *CoSENT (Cosine Sentence) Loss* (Jianlin, 2022), while NLI continued with *Cached Multiple Negatives Ranking Loss* to preserve classification performance on classification tasks. Otherwise, we observed a noticeable performance drop on classification tasks.

- MIRACL Format: (query, doc, score (float))
- NLI Format: (sentence, paraphrase, contradiction)

For the MIRACL dataset, we used the *CoSENT Loss* (Jianlin, 2022), a ranking-based loss that emphasizes preserving the relative similarity order between sentence pairs. Given a batch of input pairs with real-valued similarity labels, the CoSENT loss computes:

$$\mathcal{L} = \log \sum_{(i,j)>(k,l)} (1 + \exp(s_{(i,j)} - s_{(k,l)}))$$

where, (i,j) and (k,l) are any pairwise examples in the batch such that the label of (i,j) is greater than that of (k,l), and $s(i,j)$ is their cosine similarity. This loss encourages the model to maintain correct ranking among sentence pairs, rather than regress to a specific value. Compared to *Cosine Similarity Loss*, anecdotal experiments and prior works suggest *CoSENT* yields a stronger training signal, faster convergence, and improved retrieval performance.

We used a learning rate of 5e-5 across all training stages. The warm-up step took approximately 20 hours, and the two fine-tuning stages required about four hours in total. All training was performed on 8 NVIDIA A100-40GB GPUs.

4.3 Tooka-SBERT-Large

We use TookaBERT-Large as the backbone for our Large model, as illustrated in Fig. 1b.

Stage 1 - Warmup on TLPC: As in the Small model, we train for one epoch on the Targoman news dataset using the *Cached Multiple Negatives Ranking Loss*.

Stage 2 – Fine-tuning on Raft + NLI + MIRACL: We trained for 1 epoch across all three datasets using proportional sampling. Different loss functions were used for different views of each dataset, as shown in Table 2. To effectively leverage diverse supervision signals and task types, we used multiple loss functions tailored to each dataset's structure and goal. *Cached Multiple Negatives Ranking Loss* was chosen for datasets like NLI and RAFT, as it enables scalable contrastive learning by using all non-matching pairs in a batch as negatives. For MIRACL, which contains relevance scores from cross-encoder models, we used *CoSENT Loss* to optimize ranking consistency based on relative pairwise order, which aligns well with retrieval tasks. Additionally, we employed *Online Contrastive Loss* on MIRACL's binary

Model	Params	Pair Classification	Classification				Average	
		FarsTail Avg. Precision	**Massive Intent** Accuracy	**Massive Scenario** Accuracy	**Multilingual Sentiment** Accuracy	**Persian Food Sentiment** Accuracy	**Pair Classification & Classification**	**Overall**
e5-base-v2	109M	57.23	29.84	33.21	58.28	58.04	47.32	33.05
e5-large-v2	335M	59.25	35.75	38.06	57.79	57.18	49.61	34.65
multilingual-e5-small	118M	71.56	57.17	62.84	75.42	74.37	68.27	67.46
multilingual-e5-base	270M	70.76	61.53	65.22	76.34	75.74	69.92	69.33
multilingual-e5-large	560M	72.55	65.31	68.76	77.47	77.16	72.25	**71.44**
LaBSE	471M	62.93	62.33	67.43	72.44	72.09	67.44	55.15
gte-multilingual-base	305M	72.65	62.29	67.88	71.84	70.90	69.11	68.28
Qwen3-Embedding-0.6B	596M	73.23	68.91	72.45	69.24	68.01	70.37	67.00
jina-embeddings-v3	572M	71.88	**72.60**	**81.88**	**81.48**	**81.11**	**77.79**	71.37
openai-text-embedding-ada-002	-	65.03	52.00	56.75	71.11	70.27	63.03	53.83
openai-text-embedding-3-small	-	68.85	51.99	57.07	66.55	65.83	62.06	54.44
openai-text-embedding-3-large	-	72.45	64.80	70.26	77.01	75.98	72.10	65.77
maux-gte-persian	305M	63.80	63.51	68.19	71.88	70.68	67.61	65.39
sentence-transformer-parsbert-fa	163M	58.92	44.13	51.84	55.74	55.95	53.32	39.41
persian-embeddings	560M	71.83	64.12	73.78	67.37	66.79	68.78	64.42
Persian_Sentence_Embedding_v3	560M	69.16	63.19	71.01	72.74	72.06	69.63	62.94
bert-zwnj-wnli-mean-tokens	118M	56.09	52.76	58.24	59.64	59.38	57.22	43.07
roberta-zwnj-wnli-mean-tokens	118M	54.98	51.41	59.53	57.65	57.11	56.13	42.02
Tooka-SBERT (Ours)	353M	**81.52**	64.39	67.59	77.17	77.01	73.54	62.54
Tooka-SBERT-V2-Small (Ours)	123M	75.69	65.33	69.23	77.51	76.56	72.86	69.49
Tooka-SBERT-V2-Large (Ours)	353M	80.24	67.87	72.70	79.38	78.97	75.83	70.54

Table 3: Performance on Pair Classification and Classification tasks.

relevance data to directly optimize embedding separation between relevant and irrelevant pairs. For NLI's 3-way labeled format (entailment, contradiction, neutral), we applied *Softmax Loss*, a classification-based loss that encourages distinct clustering of semantic classes in the embedding space. This diverse loss setup enabled us to train a general-purpose model capable of strong performance across different tasks.

We used a learning rate of 5e-5 for the warm-up phase and 1e-5 for fine-tuning. The warm-up step took approximately 26 hours, while fine-tuning required about three hours. All training was conducted on 8 NVIDIA A100-40GB GPUs.

5 Evaluations

We evaluated our models on PTEB (**P**ersian **T**ext **E**mbedding **B**enchmark), which we created by selecting and unifying the Persian-language tasks available in the MTEB suite (Muennighoff et al., 2022) and enhancing key evaluation protocols to ensure fair and rigorous assessment. While PTEB utilizes the original MTEB evaluation code for most tasks, we implemented a critical correction for the MIRACLReranking task. PTEB includes evaluation on retrieval, classification, pair-classification and reranking, offering a comprehensive assessment of sentence embeddings in Persian.

5.1 Modification to the MIRACLReranking Protocol

For the Persian MIRACLReranking task, we identified a significant issue in the original MTEB benchmark's evaluation script that could lead to an inaccurate assessment of model performance. The standard protocol evaluates a model's ability to rerank a list of 100 candidate documents for each of the 632 queries. The primary evaluation metric is the Normalized Discounted Cumulative Gain (nDCG). However, we found that for certain queries, the provided set of 100 candidates did not contain any of the ground-truth positive documents. This setup flaw means that even a perfect model would score an nDCG of 0 on these samples, as it's impossible to rank documents that are not present in the candidate pool.

To ensure a more fair and rigorous evaluation, we implemented the following modifications to the evaluation code for each query:

Injecting Positive Documents: We

Model	Params	Reranking		Retrieval			Average	
		MIRACL nDCG@10	Wikipedia MAP	NeuCLIR2023 nDCG@20	MIRACL nDCG@10	Wikipedia nDCG@10	Retrieval & Reranking	Overall
e5-base-v2	109M	11.38	60.94	1.89	0.26	19.42	18.78	33.05
e5-large-v2	335M	14.50	63.50	2.05	0.16	18.31	19.70	34.65
multilingual-e5-small	118M	61.57	86.80	43.63	53.34	87.87	66.64	67.46
multilingual-e5-base	270M	65.23	86.78	46.10	57.48	88.11	68.74	69.33
multilingual-e5-large	560M	**67.72**	**89.32**	46.67	**59.01**	**90.40**	**70.62**	**71.44**
LaBSE	471M	32.79	82.42	21.52	10.53	67.06	42.86	55.15
gte-multilingual-base	305M	63.11	84.38	50.94	53.89	84.94	67.45	68.28
Qwen3-Embedding-0.6B	596M	61.20	87.33	42.30	40.60	86.78	63.64	67.00
jina-embeddings-v3	572M	49.67	79.58	**51.36**	55.15	89.04	64.96	71.37
openai-text-embedding-ada-002	-	37.16	84.41	15.79	17.29	72.77	45.48	54.26
openai-text-embedding-3-small	-	38.62	80.93	20.33	22.84	75.25	47.59	54.83
openai-text-embedding-3-large	-	54.20	85.22	39.44	39.27	85.11	60.65	66.37
maux-gte-persian	305M	61.77	80.61	44.22	50.80	78.45	63.17	65.39
sentence-transformer-parsbert-fa	163M	21.84	61.47	6.61	1.95	35.65	25.50	39.41
persian-embeddings	560M	51.89	83.47	44.16	37.11	83.71	60.07	64.42
Persian_Sentence_Embedding_v3	560M	48.26	82.62	35.07	33.39	81.93	56.25	62.94
bert-zwnj-wnli-mean-tokens	118M	20.66	73.28	5.03	4.35	41.29	28.92	43.07
roberta-zwnj-wnli-mean-tokens	118M	20.49	72.11	5.27	4.34	37.34	27.91	42.02
Tooka-SBERT (Ours)	353M	40.16	80.71	36.48	21.32	79.02	51.54	62.54
Tooka-SBERT-V2-Small (Ours)	123M	61.50	85.30	47.80	50.24	85.69	66.11	69.49
Tooka-SBERT-V2-Large (Ours)	353M	60.09	86.78	47.19	44.67	87.53	65.25	70.54

Table 4: Performance on Reranking and Retrieval tasks.

augment the candidate list by adding all ground-truth positive documents associated with the query. This guarantees that all relevant documents are available to be ranked.

Adding Negative Documents: To increase the task's difficulty, we also incorporate the provided negative documents into the candidate list, challenging the model to distinguish between relevant and highly similar irrelevant documents.

Deduplication: Finally, we process the augmented candidate list to remove any duplicate documents. This step, implemented using set operations, cleans the data and ensures each unique document is considered only once in the ranking process.

5.2 Evaluated Models

To establish a comprehensive comparison, we evaluated a wide range of state-of-the-art text embedding models. Our evaluation includes prominent open-source multilingual models such as the E5 series (Wang et al., 2022, 2024), LaBSE (Feng et al., 2022), GTE (Zhang et al., 2024), Qwen3-Embedding (Zhang et al., 2025), and Jina Embeddings v3 (Sturua et al., 2025). We also benchmark against widely-used proprietary models from OpenAI, including text-embedding-ada-002, text-embedding-3-small, and text-embedding-3-large (Neelakantan et al., 2022).

Furthermore, to establish strong language-specific baselines, we assess several models explicitly trained or fine-tuned for Persian. These include maux-gte-persian (Mirzaei, 2024), models based on ParsBERT (Ahmadi, 2024), and other community-driven efforts like persian-embeddings (Heydari, 2024), Persian Sentence Embedding v3 (Sobhi, 2024), and sentence transformers derived from Zwnj models (Farahani, 2020). We compare the performance of these established models against our proposed models: Tooka-SBERT, Tooka-SBERT-V2-Small, and Tooka-SBERT-V2-Large.

6 Results

Table 3 presents the evaluation results on pair classification and classification tasks, while Table 4 reports performance on retrieval and reranking tasks. The results compare Tooka-SBERT against state-of-the-art embedding models. Among all models, Tooka-SBERT-V2-Large ranked third overall with an average score of 70.54%, showing strong performance in pair classification (80.24%) and consistent scores across reranking and classification tasks. Tooka-SBERT-V2-Small, while more compact, also demonstrated competitive results with an average of 69.49%, outperforming several larger models such as multilingual-e5-base (69.33%) and Qwen3-Embedding-0.6B (67.00%). The original Tooka-SBERT model achieved the highest pair classification score (81.52%) but lagged in reranking and retrieval tasks, suggesting improvements in V2 versions enhanced generalization across task types. Compared to the baselines, both V2 models consistently ranked in the top 5 across most tasks, confirming the effectiveness of our training strategy on Persian-specific data.

7 Conclusion

In this work, we presented Tooka-SBERT, a lightweight yet competitive Persian sentence embedding model aimed at improving semantic understanding in low-resource settings. Through a combination of pretraining on Persian news data and fine-tuning on synthetic sentence pairs, Tooka-SBERT achieves strong performance on the PTEB benchmark, surpassing widely-used multilingual baselines. Our models strike a balance between effectiveness and efficiency, making them practical for real-world applications on both GPU and CPU.

Limitations

While Tooka-SBERT achieves strong performance across various Persian tasks, it has several limitations. First, it is specifically designed for Persian and does not generalize to multilingual settings. Second, due to the scarcity of high-quality Persian datasets, we relied on synthetic data generation, which may introduce biases. Third, both the small and large variants have a relatively small parameter count and context window, which may limit performance on complex or long-context tasks compared to larger-scale models.

References

Amir Masoud Ahmadi. 2024. sentence-transformer-parsbert-fa. https://huggingface.co/myrkur/sentence-transformer-parsbert-fa. Sentence Transformer model fine-tuned from HooshvareLab/bert-base-parsbert-uncased.

Hossein Amirkhani, Mohammad AzariJafari, Soroush Faridan-Jahromi, Zeinab Kouhkan, Zohreh Pourjafari, and Azadeh Amirak. 2023. Farstail: a persian natural language inference dataset. *Soft Computing*.

David Blei, Andrew Ng, and Michael Jordan. 2001. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Kasra Darvishi, Newsha Shahbodaghkhan, Zahra Abbasiantaeb, and Saeedeh Momtazi. 2023. Pquad: A persian question answering dataset. *Computer Speech & Language*, 80:101486.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the Association for Information Science and Technology*, 41(6):391–407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Liat Ein Dor, Yosi Mass, Alon Halfon, Elad Venezian, Ilya Shnayderman, Ranit Aharonov, and Noam Slonim. 2018. Learning thematic similarity metric from article sections using triplet networks. In *Proceedings of the 56th Annual Meeting of the Association for*

Computational Linguistics (Volume 2: Short Papers), pages 49–54, Melbourne, Australia. Association for Computational Linguistics.

Mehrdad Farahani. 2020. Persian - sentence transformers.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2020. Parsbert: Transformer-based model for persian language understanding. *ArXiv*, abs/2005.12515.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

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

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mohammad Hassan Heydari. 2024. persian-embeddings. `https://huggingface.co/heydariAI/persian-embeddings`. XLM-RoBERTa-base model fine-tuned on a large Persian corpus for contextual embeddings.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Su Jianlin. 2022. Cosent (i): A more effective sentence embedding scheme than sentence-bert.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabagdi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, and 6 others. 2021. ParsiNLU: A suite of language understanding challenges for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and 1 others. 2022. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. Cite arxiv:1907.11692.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Mani Mirzaei. 2024. maux-gte-persian. `https://huggingface.co/xmanii/maux-gte-persian`. Sentence-Transformer model fine-tuned from Alibaba-NLP/gte-multilingual-base using GPT-4 translated Persian sentence pairs.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Such, Kenny Hsu, and Lilian Weng. 2022. Text and code embeddings by contrastive pre-training.

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

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

MohammadAli SadraeiJavaheri, Ali Moghaddaszadeh, Milad Molazadeh, Fariba Naeiji, Farnaz Aghababaloo, Hamideh Rafiee, Zahra Amirmahani, Tohid Abedini, Fatemeh Zahra Sheikhi, and Amirmohammad Salehoof. 2024. Tookabert: A step forward for persian nlu. *arXiv preprint arXiv:2407.16382*.

Mehran Sarmadi, Morteza Alikhani, Erfan Zinvandi, and Zahra Pourbahman. 2025. Hakim: Farsi text embedding model. *Preprint*, arXiv:2505.08435.

Mohamad Sobhi. 2024. Persian_sentence_embedding_v3. `https://huggingface.co/Msobhi/Persian_Sentence_Embedding_v3`. Sentence Transformer model fine-tuned from XLM-RoBERTa-large on Persian datasets like ParsiNLU and PQuAD.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2025. Jina embeddings v3: Multilingual text encoder with low-rank adaptations. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part V*, page 123–129, Berlin, Heidelberg. Springer-Verlag.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Targoman. 2022. Persian web scraper. `https://github.com/Targoman/PersianWebScraper`.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *ArXiv*, abs/2212.03533.

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

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 538–548, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, and 1 others. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

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