Journal of Information Systems and Telecommunication
Vol.13, No.1, January-March 2025, 24-32
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

A Turkish Dataset and BERTurk-Contrastive Model for Semantic Textual Similarity

Somaiyeh Dehghan[1*], Mehmet Fatih Amasyali[1]

[1].Department of Computer Engineering, Yildiz Technical University, Istanbul 34220, Turkey

Received: 27 Sep 2024/ Revised: 04 Apr 2025/ Accepted: 04 May 2025

Abstract

Semantic Textual Similarity (STS) is an important NLP task that measures the degree of semantic equivalence between two texts, even if the sentence pairs contain different words. While extensively studied in English, STS has received limited attention in Turkish. This study introduces BERTurk-contrastive, a novel BERT-based model leveraging contrastive learning to enhance the STS task in Turkish. Our model aims to learn representations by bringing similar sentences closer together in the embedding space while pushing dissimilar ones farther apart. To support this task, we release SICK-tr, a new STS dataset in Turkish, created by translating the English SICK dataset. We evaluate our model on STSb-tr and SICK-tr, achieving a significant improvement of 5.92 points over previous models. These results establish BERTurk-contrastive as a robust solution for STS in Turkish and provide a new benchmark for future research.

Keywords: Semantic Textual Similarity; Contrastive Learning; Deep Learning; BERT; BERTurk; Turkish Language.

1- Introduction

Semantic Textual Similarity (STS) is a fundamental task in NLP that aims to measure the similarity of the semantic meaning of given texts. STS has a crucial role in various NLP downstream tasks, including information retrieval, text summarization, text classification, sentiment analysis, question answering, machine translation, automatic essay scoring, named entity recognition, plagiarism check, and many more. Many methods have been proposed for measuring STS including traditional methods (e.g., BOW and TF-IDF), neural embedding models (e.g., Word2Vec [1] and GloVe [2]), and deep contextualized language models (e.g., BERT [3]).

Traditional STS measurement methods only focus on a lexical level and do not consider the semantic information of words [4, 5]. For example, the two sentences “How old are you?” and “What is your age?” are completely similar in terms of meaning, but they do not have a word in common. Neural embedding-based STS measurement methods produce context-independent embeddings [6, 7]. While the meaning of words can change according to their context. For example, in these two sentences “I open a bank account.” and “The Ahilya fort on the banks of the river Narmada is amazing to see.”, the word *bank* has completely different meanings.

Recent methods of measuring STS have been able to overcome these weaknesses using deep contextualized embedding models. BERT [3] is a language model whose main technical innovation is the use of Transformers. The Transformer-based architecture of BERT uses the amazing attention mechanism that learns contextual relationships between words in a sequence of text. Moreover, BERT supports transfer learning and fine-tuning for specific tasks like STS. BERT has proven to be highly successful in a variety of NLP tasks, such as sentiment analysis [8], text classification [9], text chunking [10], and hate speech detection [11, 12, 13], demonstrating its versatility and effectiveness across different domains.

In this study, we propose a BERTurk model using contrastive learning for Semantic Textual Similarity. Our model seeks to learn a embedding space in which pairs of similar sentences remain close to each other while dissimilar sentence pairs are pushed apart. In addition, we also prepare an STS dataset for Turkish, namely SICK-tr. We evaluate our model on two Turkish STS benchmarks, STSb-tr [14] and our SICK-tr dataset. The evaluation findings show that our model performs noticeably better than previous models, demonstrating superior accuracy in capturing semantic similarities in Turkish texts, and setting a new standard for STS tasks in this language.

The proposed model and released dataset are available in our GitHub repository[1].

The current study provides significant contributions by attempting to fill several gaps as follows:

- First, the study extends the limited research on the STS task in the Turkish language, addressing a critical need in NLP for low resource languages.
- Second, the study is the first to consider contrastive learning for the STS task in Turkish, so that this method not only improves the precision of semantic similarity assessments but also sets a precedent for future research to use contrastive learning techniques in other low resource languages.
- Third, the study significantly expands the limited STS benchmarks in Turkish by releasing the SICK-tr dataset. This new dataset serves as a valuable resource for the NLP community, providing a robust foundation for future research and development in STS tasks for the Turkish language.

The remainder of the paper is structured as follows: A brief overview of related work is given in Section 2. The methodology for preparing SICK-tr dataset and our proposed model, BERTurk-contrastive, is provided in Section 3. The experiments are described in Section 4. The final portion includes conclusions and information about future work.

2- Related Work

There are many studies focusing on STS in other languages. However, to the best of our knowledge, there have been few studies in the literature for measuring semantic similarity of Turkish texts. In addition, there are seven standard benchmarks for evaluating STS in English, including STS12-STS16 [15-19], STSb [20], and SICK [21], while the only Turkish STS dataset is STSb-tr [14], which was created in 2021 by translating STSb using Google Cloud Translation API.

Ref. [14] proposed a BERT-based model for semantic textual similarity. They fine-tuned BERTurk using Cross-Entropy (CE) and Mean Squared Error (MSE) objectives on the NLI-tr [22] and STSb-tr [14] datasets, respectively. They achieved a Spearman's rank correlation of 83.31% on the STSb-tr test set for the S-BERTurk model. Ref. [23] proposed a statistical method for semantic textual similarity in Turkish news using Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA). They were able to predict the similarity between two news articles. However, the news articles they used in the experiments were few and had many words in common.

In recent years, there has been an increasing amount of literature on contrastive learning for fine-tuning BERT on semantic similarity in the English language. Contrastive learning is a deep metric learning method that encourages a model to learn an embedded space in which similar (positive) data samples (x_i, x_i^+) remain close to each other, while dissimilar (negative) data samples (x_i, x_i^-) are further apart.

Ref. [24] proposed $\text{SimCSE}_{\text{unsup}}$ and $\text{SimCSE}_{\text{sup}}$ models using self-supervised and supervised contrastive learning, respectively, to fine-tune BERT. They achieved the best results in the supervised setting with an average Spearman's rank correlation of 81.57% on seven standard STS benchmarks in English.

Ref. [25] proposed a supervised multiple positives and negatives contrastive learning model, SupMPN, to fine-tune BERT. Their idea was that by using multiple positives (similar sentences), the model would generalize in such a way that it could simultaneously bring together similar sentences in the embedding space, and by using multiple negatives (dissimilar sentences), the model would generalize to improve the distinction between similar and dissimilar sentences. They achieved an average Spearman's rank correlation of 82.07% on seven standard STS benchmarks in English.

Ref. [26] proposed a curriculum contrastive learning model (SelfCCL) by transferring self-taught knowledge for fine-tuning BERT, which mimics the human learning process. Their model learns by contrasting similar and dissimilar sentences, starting from the simplest to the hardest triplets(x_i, x_i^+, x_i^-). They achieved an average Spearman's rank correlation of 81.80% on seven standard STS benchmarks in English.

3- Methodology

This section first describes the preparation process of the SICK-tr dataset, followed by an introduction to our proposed model, BERTurk-contrastive.

3-1- Providing SICK-tr Dataset

SICK [21], an acronym for Sentences Involving Compositional Knowledge, contains about 10,000 sentence pairs with a wealth of lexical, syntactic, and semantic phenomena. Each pair of sentences has two types of annotations: relatedness and entailment. The human

[1] Our pre-trained model and released dataset are publicly available at:
https://github.com/SoDehghan/BERTurk-contrastive
https://github.com/SoDehghan/SICK-TR

relatedness score ranges from 1 to 5, and there are three categories of entailment relations: entailment, contradiction, and neutral.

Table 1. Some Translation Examples by Google Translation API in SICK-tr Dataset

Sentence 1	Sentence 2	Relatedness Score	Relationship
Bir kadın bir makineyle dikiyor. (A woman is sewing with a machine.)	Bir kadın dikiş için yapılmış bir makine kullanıyor. (A woman is using a machine made for sewing.)	4.8	Gereklilik (Entailment)
Genç çocuklar bir parkta yeşil bir futbol topu ile poz veriyor. (The young boys are posing with a green soccer ball in a park.)	Bir topun önünde dört erkek yan yana diz çöküyor. (Four boys are kneeling next to each other in front of a ball.)	3.5	Nötr (Neutral)
Kameralı bir adam konuyu inceliyor. (A man with a camera is studying the subject.)	Konuyu inceleyen kameralı bir insan yok. (There is no man with a camera studying the subject.)	3.6	Çelişki (Contradiction)

Table 2. SICK and SICK-tr statistics

Dataset	Size of Vocabulary	Average Word Length	Average Sentence Length
SICK [21]	2,551	6.38	9.65
SICK-tr (ours)	4,484	7.31	6.79

Table 3. Example of some Translation Errors from English to Turkish for SICK-tr (ours)

	Error Type	English Sentence	Turkish Translation Using Google Translation API	Corrected Turkish Translation
1	Sentiment	A skilled person is riding a bicycle on one wheel.	Yetenekli bir kişi bir tekerleğe bisiklet sürüyor.	Yetenekli bir kişi tek tekerlek üzerinde bisiklet sürüyor.
2	Syntax	A brown dog is attacking another animal in front of the man in pants.	Kahverengi bir köpek, pantolondaki adamın önünde başka bir hayvana saldırıyor.	Kahverengi bir köpek, pantolonlu adamın önünde başka bir hayvana saldırıyor.

We use a variant of SICK that is located in the SentEval GitHub repository [27]. The train-split has 4,500 pairs, the development-split has 500 pairs, and the test-split has 4,927 pairs.

We translated the English SICK dataset using Google Cloud Translation API[2], creating a variant called SICK-tr, and released it in our GitHub repository. The translation quality and adherence to the original labels have not been verified by human experts. Table 1 shows some sentence pairs from SICK-tr translated by Google Translation API, and Table 2 shows some statistics on word and sentence lengths in both SICK and SICK-tr datasets.

3-1-1- Error Types in Translation from English to Turkish

According to a study conducted on translation from Turkish to English using Google Translation [28], there are five major types of errors, including lexical errors, syntactic errors, semantic errors, morphological errors, and pragmatic errors in machine translation. Although their study focused on Turkish to English translation, we also observed the same errors in the translation from English to Turkish. However, we have not changed them, as they are few in number, and generally, such translation errors are not considered to be a major problem in STS [14]. Table 3 shows some examples of these errors. As shown in Table 3, in example 2, the preposition “in” in “man in pants” means “with” in English. However, it was translated as if it meant “içinde” (inside) in Turkish.

[2] https://cloud.google.com/translate

Table 4: Training Setting for our Models. CE: Cross Entropy, SCL: Supervised Contrastive Loss, MSE: Mean Squared Error

Model	Training Dataset	Objective Function	Batch Size	Training Epochs
S-BERTurk-nli-ce [14] (reproduced)	NLI-tr	CE	512	6
S-BERTurk-nli-contrastive (ours)	NLI-tr	SCL	512	6
S-BERTurk-nli-stsb-contrastive-mse (ours)	NLI-tr, STSb (train-split)	SCL, MSE	512, 256	6, 8

3-2- BERTurk-Contrastive Model

Contrastive learning is a type of self-supervised learning approach used to learn representations of data by contrasting positive pairs (anchor-positive: similar or related data points) against negative pairs (anchor-negative: dissimilar or unrelated data points). Figure 1 shows contrastive learning idea.

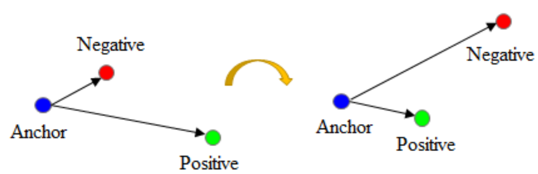

Fig 1. Contrastive learning Idea [30]

We employ the supervised contrastive loss from [26], which incorporates a hard negative to develop a version of the NT-Xent loss [29]. In a mini-batch, the Supervised Contrastive Loss (SCL) for a triplet in the form anchor-positive-negative (x_i, x_i^+, x_i^-) is given as follows:

$$L_{SCL} = -log \frac{e^{(sim(x_i, x_i^+)/\tau)}}{\sum_{j=1}^{N}\left(e^{(sim(x_i, x_j^+)/\tau)} + e^{(sim(x_i, x_j^-)/\tau)}\right)} \quad (1)$$

where $sim(\cdot)$ is the standard cosine similarity, and τ is a temperature parameter to scale the cosine similarity.

4- Experiments

4-1- Training Dataset

To train our model, we employ the Natural Language Inference (NLI) dataset in Turkish (NLI-tr) [22]. NLI is the process of determining, given a premise, whether a hypothesis is true (entailment), false (contradiction), or indeterminate (neutral). NLI-tr is a collection of two large datasets that were created by translating the SNLI [31] and MultiNLI [32] fundamental NLI corpora using Amazon Translate.

Our model's inputs are triplets in the form of (x_i, x_i^+, x_i^-), where entailment hypotheses are treated as positives and contradiction hypotheses are as negatives for the premise sentence (anchor). That is, we use only the entailment and contradiction labels, ignoring the neutral labels. Our training dataset contains roughly 300K input triplets in total.

4-2- Training Setup

The Hugging Face Model Hub hosts a pre-trained BERTurk model as our starting point. We employ the Sentence-BERT bi-encoder architecture of Sentence Transformers as described by [33]. We reproduced the S-BERTurk-nli-ce model based on [14], which was trained on the NLI-tr dataset as a three-way classification problem (entailment, contradiction, and neutral) using cross-entropy (CE) loss.

We trained two models, S-BERTurk-nli-contrastive and S-BERTurk-nli-stsb-contrastive-mse models. For S-BERTurk-nli-contrastive model, we trained BERTurk on NLI-tr using SCL. For the S-BERTurk-nli-stsb-contrastive-mse model, we first trained BERTurk on NLI-tr using SCL and then fine-tuned it on STSb-tr (train-split) using MSE (Mean Squared Error) loss.

The STSb-tr dataset, like the SICK-tr dataset, contains pairs of sentences whose degree of similarity is annotated in the range between 0 and 5. So in this case (a regression problem), MSE loss is used to compute the cosine similarity score between sentence pairs as follows:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (2)$$

where y_i and $\hat{y}_i$ are the desired values and predicted values, respectively. We have summarized the information about the training settings for our reproduced and proposed models in Table 4.

Table 5. Results of the two Turkish STS Benchmark Evaluations. For each Benchmark, a Spearman's Rank Correlation as $\rho \times 100$ is Provided in the Columns. The best Results are in Bold for Each Column.

Model	Objective Function	STSb (test-split)	SICK-tr (test-split)	Average
No fine-tuned has been done				
BERTurk (baseline model)	-	55.23	55.67	55.45
Only trained on NLI-tr				
S-BERTurk-nli-ce [14] (reproduced)	CE	72.74	70.21	71.47
S-BERTurk-nli-contrastive (ours)	SCL	**78.43**	**76.35**	**77.39**
First trained on NLI-tr and then fined-tuned on STSb-tr (train-split)				
S-BERTurk-nli-stsb-ce-mse [14] (reproduced)	CE, MSE	83.31	-	-
S-BERTurk-nli-stsb-contrastive-mse (ours)	SCL, MSE	**84.38**	**76.71**	**80.51**

4-3- Evaluation on Turkish STS Benchmarks

In this experiment, we evaluate our models on the STSb-tr (test-split) and SICK-tr (test-split) datasets. We compare our proposed models to BERTurk (the baseline model), S-BERTurk-nli-ce [14] (our reproduced model), and S-BERTurk-nli-stsb-ce-mse model [14]. Table 5 shows the results.

Results: As seen in Table 5, our models outperform the previous models, demonstrating significant advancements in accuracy and in efficiency. S-BERTurk-nli-contrastive model achieved an average improvement of 5.92 points (71.47% vs. 77.39%) compared to S-BERTurk-nli-ce (our reproduced model). Moreover, S-BERTurk-nli-stsb-contrastive-mse model achieved an improvement of 1.07 points (83.31% vs. 84.38%) on the STSb-tr dataset compared to S-BERTurk-nli-stsb-ce-mse [14]. Our findings indicate that first replacing cross-entropy loss with contrastive loss improves accuracy, as demonstrated by the S-BERTurk-nli-contrastive model's 5.92-point improvement over the S-BERTurk-nli-ce model. Additionally, using contrastive loss followed by MSE loss further enhances performance, with the S-BERTurk-nli-stsb-contrastive-mse model achieving a 1.07-point improvement (83.31% vs. 84.38%) on the STSb-tr dataset compared to the S-BERTurk-nli-stsb-ce-mse model [14].

4-4- Visualizing Sentence Embedding Space

In this experiment, we visualize the embeddings of nine sentences from SICK-tr to demonstrate the ability of our proposed model, S-BERTurk-nli-contrastive, to create a better embedding space for similar and dissimilar sentences. As explained in Section 3.1, each pair of sentences in the SICK dataset is labeled in two ways: relatedness and entailment. Therefore, we chose three anchor sentences on three different topics and their entailment and contradiction sentences as similar and dissimilar sentences, respectively, which makes nine sentences.

We use t-SNE [34], short for t-student Distributed Stochastic Neighbor Embedding, which is an unsupervised machine learning tool for visualizing high-dimensional data. t-SNE converts similarities between data points using a normal distribution in a high-dimensional space and a t-distribution in a low-dimensional space, respectively. Then, it tries to optimize the difference between the probability distributions of these two spaces using a cost function called Kullback-Leibler divergence (KL).

Figures 2, 3, and 4 show the embedding space for BERTurk (baseline model), S-BERTurk-nli-ce [14] (our reproduced model), and S-BERTurk-nli-contrastive (our proposed model), respectively.

Results: Figure 2 illustrates that BERTurk (baseline) fails to accurately differentiate between semantically distinct sentences, as evident from the close embeddings of sentences with vastly different meanings. For instance, sentences about a child playing and a brown dog playing with a toy are incorrectly grouped. This highlights the limitations of the baseline model in capturing semantic nuances. Figure 3 demonstrates that the S-BERTurk-nli-ce [14] model improves upon BERTurk (baseline) by grouping sentences with similar sentiment polarity (positive or negative).

Fig. 2. Visualizing Embedding Space for nine Sentences from SICK-tr Dataset by BERTurk (Baseline Model)

Fig. 3. Visualizing Embedding Space for nine Sentences from SICK-tr Dataset by BERTurk-nli-ce (our Reproduced Model) [14]

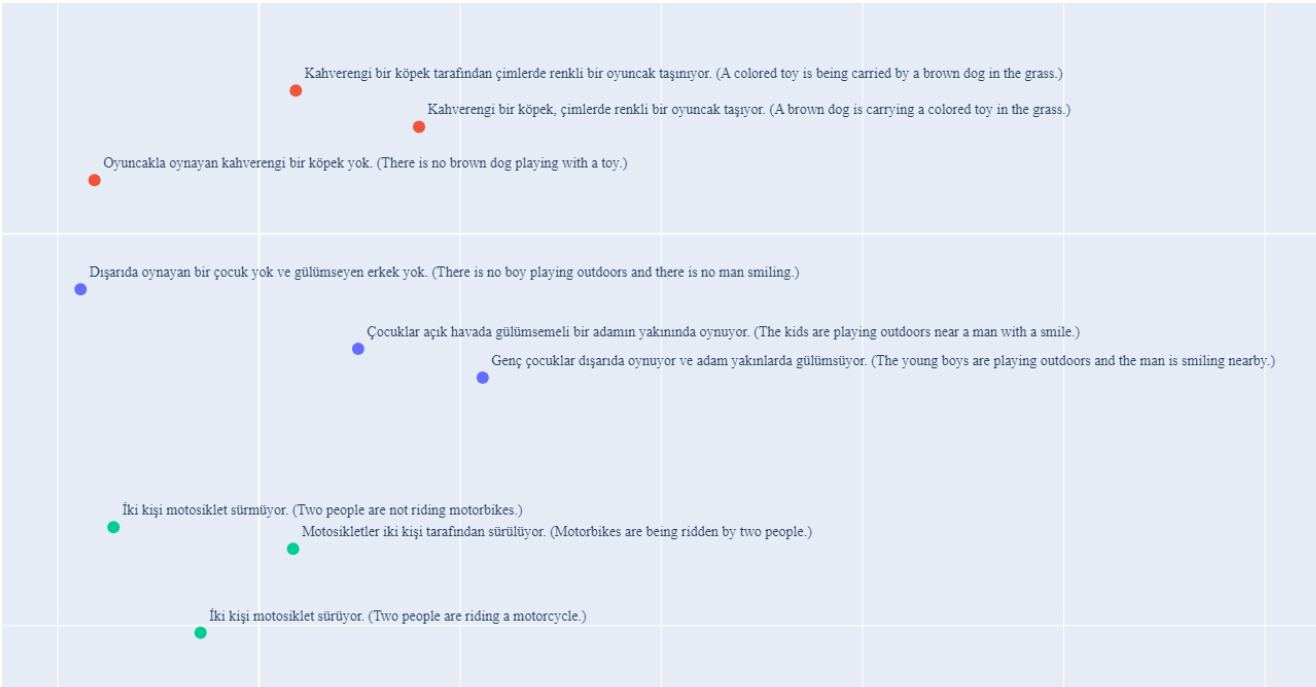

Fig. 4. Visualizing Embedding Space for nine Sentences from SICK-tr Dataset by BERTurk-nli-Contrastive (our Proposed Model)

However, it fails to capture semantic differences within the same sentiment category. For example, sentences such as “Dışarıda oynayan bir çocuk yok ve gülümseyen erkek yok. (There is no boy playing outdoors and there is no man smiling)”, “Oyuncakla oynayan kahverengi bir köpek yok. (There is no brown dog playing with a toy)”, and “İki kişi motosiklet sürmüyor. (Two people are not riding motorbikes)”, are all embedded closely due to their shared negative polarity, despite their different semantics.

Figure 4 showcases the strength of our proposed S-BERTurk-nli-contrastive model, which organizes embeddings based on both sentiment and semantics. As can be seen in figure 4, our proposed model, S-BERTurk-nli-contrastive, is able to correctly embed the sentences in the embedding space based on their concepts (topics). In addition, our proposed model is better able to distinguish between positive (similar) and negative (dissimilar) sentences for each topic. For instance, for the sentences: “Kahverengi bir köpek, çimlerde renkli bir oyuncak taşıyor. (A brown dog is carrying a colored toy in the grass.)”, “Kahverengi bir köpek tarafından çimlerde renkli bir oyuncak taşınıyor. (A colored toy is being carried by a brown dog in the grass.)”, and “Oyuncakla oynayan kahverengi bir köpek yok. (There is no brown dog playing with a toy.)”, our model successfully groups the first two sentences together due to their similar semantic meanings, both describing the action of a brown dog interacting with a toy. Meanwhile, it places the third sentence, which negates the presence of a brown dog playing with a toy, in a distinct position in the embedding space, reflecting its dissimilar semantic meaning. This demonstrates that our proposed model excels in accurately capturing both the concepts and the relationships between sentences, resulting in embeddings that align closely with their true semantic meanings.

5- Conclusion and Future Work

In this study, we proposed a BERTurk-contrastive model that used contrastive learning for the STS task in the Turkish language. This approach represents a significant advancement in the application of contrastive learning to the Turkish language, a relatively underexplored area in NLP research. Our primary contribution includes the creation of the SICK-tr dataset using the Google Translation API, which we have released publicly via GitHub for public use, providing a valuable resource and benchmark for future research on STS in Turkish.

Our evaluation results on two STS datasets, STSb-tr and SICK-tr, demonstrate that replacing cross-entropy loss with contrastive loss leads to a substantial improvement of 5.92 points (71.47% to 77.39%). This highlights the effectiveness of contrastive learning in capturing semantic similarities more accurately, particularly for low-resource languages. Additionally, visualizing the embedding space for nine sentences on three different topics shows that our model can better distinguish between similar and dissimilar sentences within each topic. This capability is crucial for enhancing the performance of various downstream NLP applications, such as text clustering, information retrieval, and question answering.

The creation of the SICK-tr dataset, coupled with the improved performance of our contrastive model, establishes a foundation for further advancements in Turkish STS tasks. Future work will extend this research by exploring state-of-the-art large language models, such as GPT, and T5, and XLM-R, alongside novel contrastive learning strategies. These efforts aim to further advance the performance and applicability of STS systems in Turkish and other low-resource languages.

Limitations

The main limitations of our work include the reliance on a translation-based dataset (SICK-tr), which may not fully capture the nuances of Turkish language structure and idiomatic expressions, potentially introducing bias. Additionally, our model is evaluated only on two datasets (STSb-tr and SICK-tr), limiting its generalizability to other domains or real-world applications. Lastly, the computational requirements of training the model may pose challenges for broader accessibility.

References

[1] T. Mikolov, K. Chen, G. S. Corrado, J. Dean, “Efficient estimation of word representations in vector space,” In Proceedings of the 2013 International Conference on Learning Representations, 2013.

[2] J. Pennington, R. Socher, C. Manning, “Glove: Global vectors for word representation,” In Proceedings of the 2014 Conference on Empirical Methods in NLP (EMNLP), pp. 1532–1543. 2014.

[3] J. Devlin, JM.-W. Chang, K. Lee, K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, In Proceedings of the 2019 Conference of the American Chapter of the Association for Computational Linguistics, Vol. 1, pp. 4171—4186, 2019.

[4] H. Cheng, S. Yat, "A Text Similarity Measurement Combining Word Semantic Information with TF-IDF Method", Chinese Journal of Computers, 2011.

[5] S. Albitar, S. Fournier, B. Espinasse, "An Effective TF/IDF-based Text-to-Text Semantic Similarity Measure for Text Classification", Web Information Systems Engineering, pp. 105-114, 2014.

[6] J. Chandra, A. Santhanam, A. Joseph, "Artificial Intelligence based Semantic Text Similarity for RAP Lyrics," 2020 International Conference on Emerging Trends in Information Technology and Engineering, pp. 1-5, 2020.

[7] E. Hindocha, V. Yazhiny, A. Arunkumar, P. Boobalan, "Short-text Semantic Similarity using GloVe word embedding", International Research Journal of Engineering and Technology (IRJET), Volume: 06, Issue: 04, Apr 2019.

[8] S. Chakraborty, “An Efficient Sentiment Analysis Model for Crime Articles’ Comments using a Fine-tuned BERT Deep Architecture and Pre-Processing Techniques”, Journal of Information Systems and Telecommunication (JIST), Vol. 45, pp. 1-11, 2024.

[9] J. Nagesh, “Hierarchical Weighted Framework for Emotional Distress Detection using Personalized Affective Cues,” Journal of Information Systems and Telecommunication (JIST), Vol. 38, pp. 89-101, 2022

[10] P. Kavehzadeh, “Deep Transformer-based Representation for Text Chunking”, Journal of Information Systems and Telecommunication (JIST), Vol. 43, pp. 176-184, 2023.

[11] S. Dehghan, B. Yanıkoğlu, “Evaluating ChatGPT’s Ability to Detect Hate Speech in Turkish Tweets,” In Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 54–59, St. Julians, Malta. Association for Computational Linguistics, 2024.

[12] S. Dehghan, B. Yanıkoğlu, “Multi-domain Hate Speech Detection Using Dual Contrastive Learning and Paralinguistic Features,” In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 11745–11755, Torino, Italia. ELRA and ICCL, 2024.

[13] S. Dehghan, M. U. Şen, B. Yanıkoğlu, “Dealing with annotator disagreement in hate speech classification,” Preprint, arXiv:2502.08266, 2025.

[14] F. B. Fikri, K. Oflazer, B. Yanıkoğlu, “Anlamsal Benzerlik için Türkçe Veri Kümesi (Turkish Dataset for Semantic Similarity)”, In Proceedings of the 29th IEEE Conference on Signal Processing and Communications Applications, Istanbul, Turkey, 2021.

[15] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, “Semeval-2012 task 6: A pilot on semantic textual similarity,” In *SEM 2012: The First Joint Conference on Lexical and Computational Semantics (SemEval 2012), Association for Computational Linguistics, pp. 385–393, 2012.

[16] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, “*sem 2013 shared task: Semantic textual similarity,” in In Second Joint Conference on Lexical and Computational Semantics (*SEM), Vol. 1, pp. 32–43, 2013.

[17] E. Agirre, C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, R. Mihalcea, G. Rigau, J. Wiebe, “Semeval-2014 task 10: Multilingual semantic textual similarity,” Association for Computational Linguistics, pp. 81–91, 2014.

[18] E. Agirre, C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, R. Mihalcea, G. Rigau, L. Uria, J. Wiebe, “Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability,” Association for Computational Linguistics, pp. 252–263, 2015.

[19] E. Agirre, C. Banea, D. Cer, M. Diab,A. Gonzalez-Agirre, R. Mihalcea, G. Rigau, J. Wiebe, “Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual

evaluation," Association for Computational Linguistics, pp. 497–511, 2016.

[20] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, L. Specia, "Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation," In Proceedings of the 11th International Workshop on Semantic Evaluation, pp. 1–14, 2017.

[21] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, R. Zamparelli, "A sick cure for the evaluation of compositional distributional semantic models," in In Proceedings of the International Conference on Language Resources and Evaluation (LREC), pp. 216–223, 2014.

[22] E. Budur, R. Özçelik, T. Güngör, "Data and Representation for Turkish Natural Language Inference", Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Nov. 2020.

[23] E. Yıldıztepe, V. Uzun, "Olasılıksal Yöntemler ile Türkçe Metinlerin Anlamsal Benzerliğinin Belirlenmesi", Sinop Üniversitesi Fen Bilimleri Dergisi, Sinop Uni J Nat Sci 3 (2): 66-78, 2018.

[24] T. Gao, X. Yao, D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings", In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021.

[25] S. Dehghan, M.F. Amasyali, "SupMPN: Supervised Multiple Positives and Negatives Contrastive Learning Model for Semantic Textual Similarity", Applied Sciences, 12:9659, 2022.

[26] S. Dehghan, M.F. Amasyali, "SelfCCL: Curriculum Contrastive Learning by Transferring Self-Taught Knowledge for Fine-Tuning BERT", Applied Sciences, Vol. 13(3):1913, 2023.

[27] A. Conneau, D. Kiela, "SentEval: An evaluation toolkit for universal sentence representations" In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan, 7--12 May, 2018.

[28] B. Koçer Güldalı, K. U. İşisağ, "A comparative study on google translate: An error analysis of Turkish-to English translations in terms of the text typology of Katherina Reiss", RumeliDE Dil ve Edebiyat Araştırmaları Dergisi, 2019.

[29] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, "A simple framework for contrastive learning of visual representations," arXiv: 2002.05709, 2020.

[30] F. Schroff, D. Kalenichenko, J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering", arXiv:1503.03832, 2015

[31] S.R. Bowman, G. Angeli, C. Potts, C.D. Manning, "A large annotated corpus for learning natural language inference", In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Portugal, 2015.

[32] A. Williams, N. Nangia, S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference", In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, Vol. 1, 2018.

[33] N. Reimers, I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert networks", In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992, 2019.

[34] L.V.D. Maaten, G.E. Hinton, "Visualizing Data Using t-SNE", Journal of Machine Learning Research, 9, pp. 2579–2605, 2008.