International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 15 (2023) pp. 605-615

Received: 06 March 2023; Accepted: 25 August, 2023; Published: 11 September, 2023

Efficient Deep Pre-trained Sentence Embedding Model for Similarity Search

Khushboo Taneja[1,*], Jyoti Vashishtha[1] and Saroj Ratnoo[1]

[1] Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology, Hisar -125001, India
khushbootaneja@gmail.com, jyoti.vst@gmail.com, ratnoo.saroj@gmail.com

***Abstract*: Nowadays, transformer is a dominant mainstream architecture for various use cases in natural language processing (NLP). The deep pre-trained Bi-directional Encoder Representations from Transformers (BERT) has obtained remarkable performance in sentence-pair regression tasks such as semantic textual similarity (STS). However, it employs a cross-encoder structure for this task and produces only token embeddings, not sentence embeddings. A high-quality sentence embedding plays a vital role in the applications where we must deal with millions of sentences like clustering and semantic search. Due to its cross-encoder structure, BERT does not scale well for large datasets and causes massive computational overhead. On the other hand, Sentence-BERT (SBERT), a pre-trained Sentence Embedding Model (SEM) (also known as sentence transformer) adapted BERT using Siamese network and trained it on Natural Language Inference (NLI) data with softmax loss (SL) for learning universal sentence representation. Its bi-encoder structure significantly reduces the computational overhead for large corpora of text. This paper presents MNRLSBERT and MNRLSRoBERTa as more efficient novel variants of SBERT in combination with Multiple Negatives Ranking Loss (MNRL), a contrastive training objective function. The novel combinations are trained with only sentence pairs of NLI data having an 'entailment' label. The performance of our models is evaluated on seven standard STS tasks and compared with competitive baseline models. By introducing an alteration in the training objective of SBERT, our models have significantly outperformed many comparable SEMs.**

***Keywords*:** BERT, Sentence transformer, Sentence embedding, Semantic textual similarity, Siamese network, Multiple negatives ranking loss

I. Introduction

BERT [1] is a word embedding transformer [2] model that has achieved outstanding performance in challenging NLP tasks including text classification, question answering, text summarization and sentence-pair regression [3][4]. Sentence-pair regression is a task in which we measure the similarity of two sentences based on their context or semantics as shown in figure 1. It plays an important role in a wide range of applications such as information retrieval, semantic search, clustering, etc.

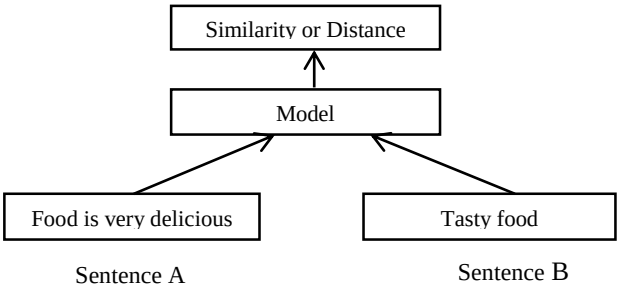


Figure 1. Sentence-pair regression

BERT uses a cross-encoder structure for the sentence-pair regression task in which pair of sentences separated by "SEP" token is processed in a single sequence, as shown in Figure 2. In this setting, both the input sentences are first converted into a sequence of tokens where a token can be a word or a piece of word and a number identifies it. This process is known as tokenization. Here, each sequence has two special tokens named "CLS" and "SEP". "CLS" indicates the beginning of the sequence, and "SEP" marks the separation between the sentences of the sequence. "PAD" tokens are appended after the "SEP" token according to the maximum sequence length of the tokenizer. These tokens are then passed to a transformer model like BERT. With this setup, BERT outputs the embedding vector for each token with 768 dimensions. However, it does not generate the embedding vector for individual sentences. BERT has achieved the most accurate results with this setup, but this approach does not scale well for large collections of text. For example, if we have a collection of 100K sentences, then to find the most similar pair, it will need 50 million inference computations which takes approximately 65 hours for execution with BERT [5]. This issue can be further alleviated if we generate and store sentence embeddings for each sentence. Researchers have found different approaches to generate sentence embeddings from the original BERT and other transformer models. The most common approach is to average the word embeddings produced by BERT. The other approach is to use the output embedding of "CLS" token which is used in classification tasks. But there is no evidence that these approaches can generate high quality sentence embeddings. It has been found that sentence embeddings produced by these approaches are

worse than averaging GloVe embeddings [5][6].

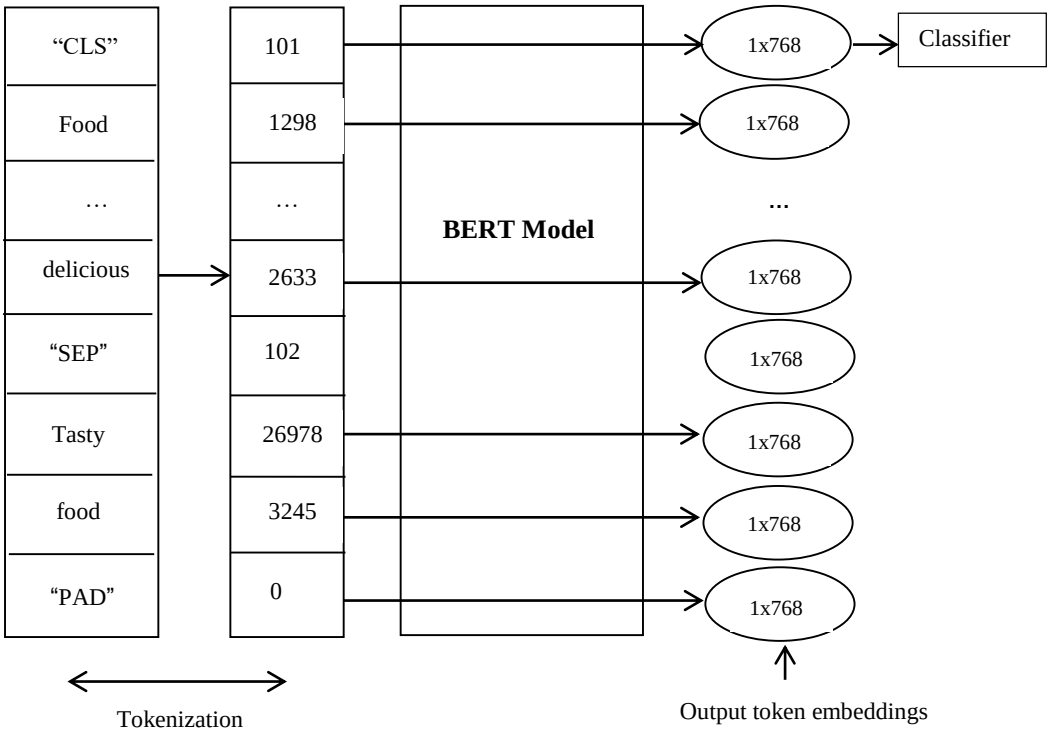


Figure 2. BERT model workflow

Reimers and Gurevych (2019) presented the first sentence transformer model known as Sentence-BERT (SBERT) [5] that can be viewed as fine-tuned version of BERT by using Siamese network [7]. With SBERT, sentence embeddings can be generated and compared with each other within a few seconds. It significantly decreased the computational burden for sizable text corpora and outperformed supervised baselines like InferSent [8] and Universal Sentence Encoder [9] for Semantic Textual Similarity (STS). SBERT laid the foundation for several subsequent advances in the field of sentence representation. Apart from it, other approaches have also been proposed in the literature to deal with inadequacy of directly using BERT embeddings for sentence representation. For example, Li et al. (2020) introduced BERT-flow [10] that transformed the BERT embeddings to alleviate anisotropy. In the same vein, Su et al. (2021) proposed BERT-whitening [11] in which the authors applied the whitening operation in traditional machine learning to enhance the isotropy and reduce the dimensionality of BERT sentence representations.

Sentence embedding (or sentence representation learning) is a growing field of research. Most of the methods in this field can be categorized as supervised and unsupervised sentence representation methods. Supervised methods like SBERT [5], InferSent [8] and Universal Sentence Encoder [9] need labelled data for training. While unsupervised sentence representation methods do not rely on labelled data. More recently, researchers have proposed various unsupervised methods that are based on contrastive learning such as IS-BERT [12], DeCLUTR [13], CT [14], SimCSE [15] and DiffCSE [16]. In these methods, they worked on different strategies to generate positive and negative examples in an unsupervised manner. Most of the existing methods for sentence representation learning are computationally expensive. Reimers and Gurevych (2019) have used softmax loss (SL) as the primary method to pre-train SBERT and SRoBERTa. However, training with softmax loss requires longer training sessions and a considerable amount of training data. Also, it does not yield optimal performance for STS and there is a scope of improvement. To overcome with this limitation, we have proposed an alternative loss function named as Multiple Negatives Ranking Loss (MNRL) to fine-tune SBERT for STS task. In this research work, we employ Siamese BERT model architecture and MNRL for building efficient pre-trained sentence embedding model. Our models, named as MNRLSBERT and MNRLSRoBERTa, have been trained on NLI [17][18] data.

MNRL is based on the principle of contrastive learning [19]. Contrastive learning is a method to learn an embedding space in which similar data pairs have closed representations and dissimilar pairs remain at distance from each other. It is a great loss function that works in the scenarios where we have positive pairs of text. For example, pairs of (query, response), pairs of (source_language, target_language), pairs of duplicate questions, etc.

In case of NLI data, the sentences with 'entailment' label provide such kind of data. For training our models with MNRL training objective, we input NLI pair of sentences in the format (anchor, positive) in which anchor entails positive. During training, the sentences having similar semantics will get embedded close to each other and dissimilar ones will stay apart in the embedding space.

We have evaluated our models on seven STS tasks including STS12 [20], STS13 [21], STS14 [22], STS15 [23], STS16 [24], STSbenchmark (STSb) [25] and the SICK-Relatedness (SICK-R) dataset [26]. The results show that our models have outperformed several SOTA supervised and

unsupervised sentence representation baseline models.

The main contribution of this research is to design optimal variants of SBERT by utilizing MNRL loss function via contrastive training objective. Optimizing a pre-trained model has a potential to improve models that are fine-tuned on it for solving downstream tasks like text classification, question/answering, information retrieval, etc.

The rest of the paper is structured as follows. Section 2 discusses the related work. Section 3 elaborates the proposed methodology. Section 4 describes the training details and section 5 presents the results obtained in the study. Finally, section 6 concludes the research work along with its future scope.

II. Related Work

Producing high quality sentence embedding is a challenging and active research area in NLP. A few unsupervised and supervised sentence representation learning methods have been proposed in this field. Supervised sentence representations use sentence pairs labels that give information about the relatedness of sentences while unsupervised sentence representations use unlabelled corpus for training.

Pre-trained word embeddings Word2Vec [27] and GloVe [6] have been effectively used for semantic representation of words and sentences. Inspired by Word2Vec, Kiros et al. (2015) presented SkipThought [28], an unsupervised method that extended the skip-gram [29] model to the sentence level. They performed the training of an encoder-decoder architecture to predict surrounding sentences. Conneau et al. (2017) proposed InferSent [8], which is siamese BiLSTM supervised network model trained on Stanford Natural Language Inference (SNLI) dataset [17] and Multi-Genre Natural Language Inference (MNLI) dataset [18]. Their work has shown the superiority of InferSent over unsupervised SkipThought. Cer et al. (2018) presented a transformer network known as Universal Sentence Encoder [9] and augmented the unsupervised learning by training on SNLI. Hill et al. (2016) [30] demonstrated that the quality of sentence embeddings is strongly influenced by the task on which they are trained. The previous work [8][9] indicates that SNLI datasets are more suitable for training SEMs. Yang et al. (2018) [31] proposed a technique for training on Reddit chats using Siamese transformer and Siamese deep averaging networks that performed well on the STSb dataset.

Devlin et al. (2019) proposed BERT [1] that has achieved cutting-edge performance on STS benchmark [25]. But the disadvantage with BERT is that it uses cross-encoder structure, and we can derive only word level embeddings and not sentence level embeddings using this setup. This results into great computation overhead to find similar sentences in a large dataset. Researchers have proposed various methods to address this limitation like utilizing the output embedding of "CLS" token or averaging the word embeddings [32][33]. But these approaches are likely to yield poor results [5]. Humeau et al. (2019) presented a method (poly-encoders) to compute a score between m context vectors and pre-computed candidate embeddings using attention to address the run-time overhead of cross-encoder BERT [34]. This approach is effective for locating the best sentence in a large corpus. For use-cases like clustering, polyencoder's computational overhead is too high and the score function is not symmetric. Reimers et. al. (2019) proposed SBERT/SRoBERTa [5] by adapting BERT/RoBERTa as underlying transformers using Siamese network to build sentence embeddings that significantly reduces the computation time over the cross-encoder BERT set up for semantic similarity. They have used the SL function to train the model. Researchers have also proposed different approaches that worked towards regularizing BERT embeddings. For example, Li et. al (2020) proposed BERT-flow [10] that transformed BERT sentence embedding to a smooth and isotropic Gaussian distribution. Also, Su et. al (2021) applied whitening operation to improve the isotropy of learned sentence representation and to reduce sentence embedding dimensionality [11].

More recently, the research trend for sentence representation is shifted towards unsupervised approaches with contrastive learning objective. Zhang et al. (2020) proposed a self-supervised learning objective based on mutual information maximization to learn semantically meaningful embeddings in an unsupervised manner [12]. Giorgi et al. (2021) presented DeCLUTR [13] which used overlapped spans to mine positive and negative examples for sentence representation with contrastive objective. Carlsson et al. (2021) introduced a self-supervised method based on contrastive tension that applied two different encoders to align the embeddings of the same sentence [14]. Gao et al. (2021) employed the drop out method to generate the positive pair of sentences for learning unsupervised sentence representations in unsupervised SimCSE and utilized annotated NLI data for supervised SimCSE [15]. Chuang et al. (2022) introduced DiffCSE [16], a method based on equivariant contrastive learning [35].

Reimers et al. (2021) implemented the idea of mapping similar sentences in different languages to the same location in vector space using the process of multilingual knowledge distillation [36]. In their work, they used MNRL function to train paraphrase model and suggested it as an optimal loss function. Tan and Koehn (2022) have fine-tuned SBERT with MNRL and used it for mining high-quality bitexts for low resource languages Khmer and Pashto [37].

Although, many approaches have been proposed in the past to generate good quality sentence embedding but still there is a scope of improvement to achieve the best scores for STS. To fill this gap, there is a need to develop models that can produce more reliable representation of sentences. In this paper, we have imbibed MNRL approach with contrastive training objective for training SBERT for producing general purpose sentence embeddings.

III. The Proposed Model

A. Siamese model architecture

We have fine-tuned SBERT [5] on NLI sentence pairs using Siamese-type model architecture [7]. Figure 3 illustrates the architecture of SBERT model. It can be viewed as a network comprised of two identical BERT models in parallel with shared weights. In reality, the architecture consists of a single BERT model in which one sentence is processed after another.

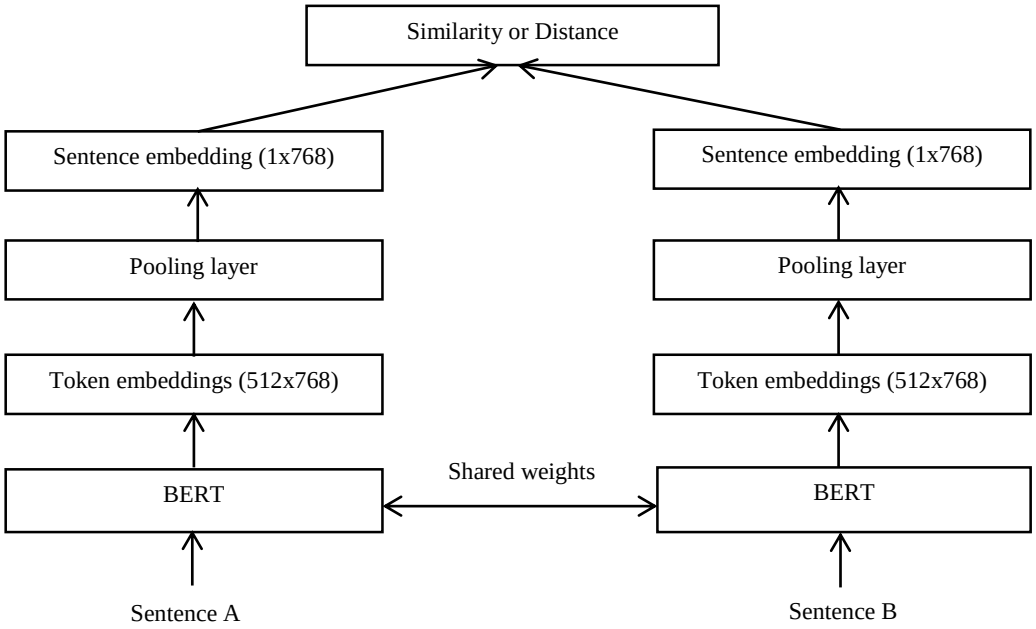


Figure 3. SBERT Siamese model architecture

In this network structure, BERT generates token level embeddings for each sentence. For example, if there are 512 tokens in a sentence, it will generate 512x768 token embeddings. SBERT adds a pooling layer on top of BERT that generates embedding vector for each sentence with 768 dimensions (1x768) using mean of all token-level embedding vector [5]. We can store this embedding and later, it can be compared with other sentence embeddings using cosine similarity only in ~0.01 seconds.

B. Loss function

The loss function plays a crucial role while fine-tuning a model on training data. The quality of sentence embedding for a specific downstream task is strongly determined by loss function. The original SBERT [5] uses SL to train model on NLI data. Using softmax loss requires a significant amount of training data and takes more training time. Additionally, it does not produce STS's best performance, and there is room for improvement. To fill this gap, we have applied an alternative loss function named as Multiple Negatives Ranking Loss (MNRL) which is described below.

1) Multiple negatives ranking loss

MNRL is a cross-entropy loss with "in-batch" negatives. The training data for MNRL approach consists of a batch of sentence pairs $[(x_1,y_1), (x_2,y_2),\ldots,(x_n,y_n)]$, where we assume that (x_i,y_i) are positive pairs and (x_i,y_j) are negative pairs for i!=j. Formally, the training objective for each batch is given as in equation (1):

$$J_{MNRL}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp \sigma(f_\theta(x_i),(f_\theta(y_i))}{\sum_{j=1}^{N} \exp \sigma(f_\theta(x_i),(f_\theta(y_j))} \quad (1)$$

Where σ is a similarity function for vectors and f_θ is the sentence encoder that embeds sentences.

During training, the distance between (x_i,y_i) is minimized and simultaneously distance between (x_i,y_j) is maximized for all i!=j as shown in figure 4. For each x_i, it considers all other y_j as negative samples i.e., for x_i, we have one positive example y_i and n-1 negative examples y_j. It then minimizes the negative log-likehood for softmax normalized scores.

We can also provide one or multiple hard negatives per (anchor, positive) pair in the triplet format: $[(x_1,y_1,n_1)]$ where n_1 is a hard negative for (x_1,y_1). The loss will use all y_j (j!=i) and all n_j as negatives for the pair (x_i,y_i).

We can use the contradiction label for NLI data to form such triplets. However, for training our models in this work, we have selected only positive sentence pairs with 'entailment' label. We have not used the hard negatives. Figure 5 illustrates the methodology and the training setup for our models.

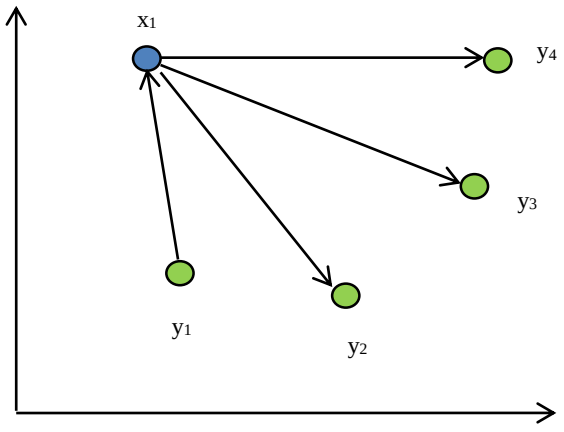


Figure 4. Distribution of similar and dissimilar sentences in the embedding space

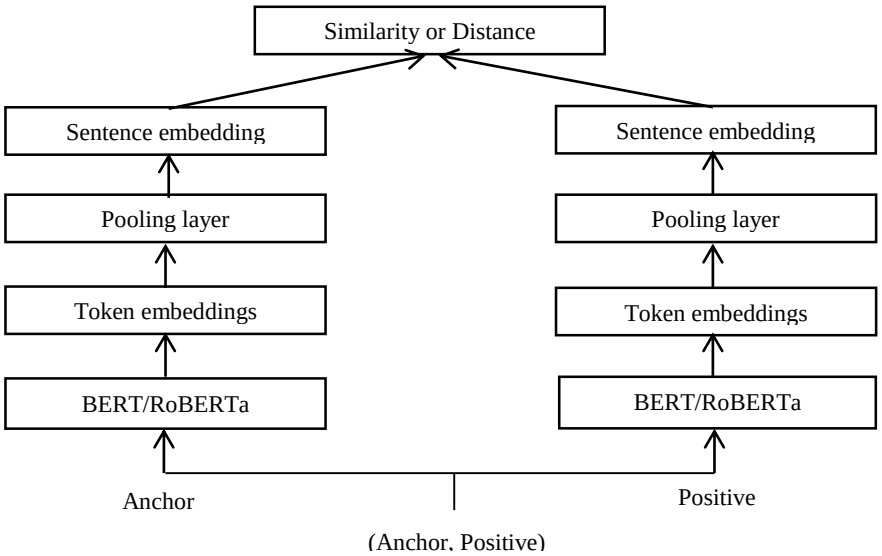


Figure 5. Methodology of the proposed approach

IV. Training Details

A pre-trained SEM requires substantial training data and fine-tuning over the target task. We have fine-tuned the transformer models adapting Siamese architecture on SNLI [17] and MNLI [18] datasets which are together named as NLI training data. We have conducted all the experiments in *Google Colab* using python programming language with single T4 GPU as hardware accelerator.

A. Dataset

SNLI consists of 5,70,000 sentence pairs and MNLI consists of 4,30,000 sentence pairs. Each sentence pair in both datasets has a premise and a hypothesis. A sentence pair is assigned *entailment, neutral or contradiction* label based on the sentence similarity. The datasets are downloaded from Hugging Face *datasets* library. SNLI consists of 550K and MNLI consists of 393K sentence pairs for training. These datasets are then merged to form training corpora of total 943K sentence pairs.

B. Data preparation

For MNRL, we have prepared the data by dropping all rows with 'neutral' and 'contradiction' label. There are some rows in both datasets with label field value -1, which means no confident class can be assigned for them. These rows are also removed using *filter* method. It reduces the training data size from 9,42,854 rows to 3,14,315 rows.

C. Model training

We have used the *sentence-transformers* library for training our models. For this, we first transformed the data into the format required by sentence-transformers library using the *InputExample* class. The *data loader* is initialized with batch size 12 (we have also tried batch size 8 and 16 but it decreased the performance) and we input the anchor-positive pairs into baseline transformer model like BERT/RoBERTa as shown in figure 6. Each model is trained for one epoch using Adam optimizer with learning rate of 2e-5. We set the linear warmup for the first 10% of the training steps. With a reduced training set, models are trained in less than 2 hours.

D. Evaluation metrics

The performance of each model is computed in terms of Spearman rank correlation, denoted by ρ , between the cosine-similarity of the sentence embeddings/vectors and manually annotated gold labels. The formula for calculating the cosine-similarity between vectors is given in equation (2)

$$Similarity(u,v) = \cos(\theta) = \frac{u.v}{\|u\|\|v\|} \quad (2)$$

Here θ is the angle between the vectors, u and v are the sentence vectors, $u.v$ calculates the dot product between u and v. $\|u\|$ and $\|v\|$ represents the magnitude of the vector.

The Spearman's rank correlation (ρ) is considered as the best metric for evaluation of STS tasks [5]. It can be calculated using formula in equation (3)

$$\rho = \frac{\sum_i (r[x_i] - r[\bar{x}])(r[y_i] - r[\bar{y}])}{\sqrt{\sum (r[x_i] - r[\bar{x}])^2 \sum_i (r[y_i] - r[\bar{y}])^2}} \quad (3)$$

Here, x_i is the ith element in the list of sentence similarity computed values, y_i is the ith element in the list corresponds to human judgement. $r[x_i]$ and $r[y_i]$ indicates the integer rank of x_i in cosine similarity vector X and integer rank of y_i in annotated similarity score vector Y respectively. $r[\bar{x}]$ and $r[\bar{y}]$ represents means of the ranks. If the correlation score evaluates near to 1, it shows that model is generating more reliable sentence embeddings.

V. Results and Discussions

A. Baseline models

We have compared our models with previous SOTA unsupervised and supervised sentence representation learning baselines on several STS tasks.

Unsupervised baselines include average GloVe embeddings [6], average BERT or RoBERTa embeddings, and post-processing methods like BERT-flow [10] and BERT-whitening [11], which focus on regularizing BERT embeddings. We have also compared our models with the most recent unsupervised contrastive learning methods such as IS-BERT [12], that maximizes the agreement between local and global features, DeCLUTR [13], which considers different spans from the same document as positive pairs, CT [14], which uses two different encoders to align the embeddings of the same sentence, SimCSE [15], which introduces the dropout noise to predict the sentence in unsupervised manner and uses NLI datasets under supervised approach, and DiffCSE [16], which is based on equivariant contrastive learning.

Supervised baselines include InferSent [8], Universal Sentence Encoder [9], and SBERT/SRoBERTa [5] along with post-processing methods BERT-flow [10], BERT-whitening [11] and CT [14] with NLI supervision setting. Our models are also compared with contrastive learning method SimCSE [15] in supervised manner.

B. Semantic textual similarity

We have performed the evaluation of MNRLSBERT and MNRLSRoBERTa on different STS tasks including the STS tasks 2012 – 2016 [8-12], STSb [25] and the SICK-R dataset [26] without using any STS specific supervision. Each dataset consists of sentence pairs which are assigned a label whose value lies between 0 and 5 based on the semantic similarity between sentences. We normalized the label values of each dataset to a scale of 0 to 1 and used sentence transformers evaluation utilities to evaluate each model.

Table 1. Performance comparison of our models with various unsupervised models in terms of Spearman rank correlation (ρ x100).

Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Average
Avg. GloVe embeddings[@]	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
BERT (first-last avg.)[#]	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
BERT-flow[#]	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
BERT-whitening[#]	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS-BERT*	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
CT-BERT[#]	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
SimCSE-BERT[#]	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
DiffCSE-BERT[$]	72.28	84.43	76.47	83.90	80.54	80.59	71.23	78.49
MNRLSBERT	72.84	81.70	74.77	82.07	79.19	83.78	75.85	**78.60**
RoBERTa(first-last avg.)[#]	40.88	58.74	49.07	65.63	61.48	58.55	61.63	56.57
RoBERTa-whitening#	46.99	63.24	57.23	71.36	68.99	61.36	62.91	61.73
DeCLUTR-RoBERTa[#]	52.41	75.19	65.52	77.12	78.63	72.41	68.62	69.99
SimCSE-RoBERTa[#]	70.16	81.77	73.24	81.36	80.65	80.22	68.56	76.57
DiffCSE-RoBERTa[$]	70.05	83.43	75.49	82.81	82.12	82.38	71.19	78.21
MNRLSRoBERTa	72.90	81.45	73.84	82.12	79.54	84.83	75.32	**78.57**

).@: results obtained from Reimers and Gurevych (2019) [5]; #: results obtained from Gao et al. (2021) [15]; *: results obtained from Zhang et al. (2020) [12]; $: results obtained from Chuang et al. (2022) [16]

Table 2. Performance comparison of our models with various supervised models in terms of Spearman rank correlation (ρ x100).

Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Average
InferSent-Glove[@]	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder[@]	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
SBERT[@]	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT-flow[#]	69.78	77.27	74.35	82.01	77.46	79.12	76.21	76.60
SBERT-whitening[#]	69.65	77.57	74.66	82.27	78.39	79.52	76.91	77.00
SimCSE-BERT[#]	75.30	84.67	80.19	85.40	80.82	84.25	80.39	**81.57**
MNRLSBERT	72.84	81.70	74.77	82.07	79.19	83.78	75.85	**78.60**
SRoBERTa[@]	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
SRoBERTa-whitening[#]	70.46	77.07	74.46	81.64	76.43	79.49	76.65	76.60
SimCSE-RoBERTa[#]	76.53	85.21	80.95	86.03	82.57	85.83	80.50	**82.52**
MNRLSRoBERTa	72.90	81.45	73.84	82.12	79.54	84.83	75.32	**78.57**

@: results obtained from Reimers and Gurevych (2019) [5]; #: results obtained from Gao et al. (2021) [15].

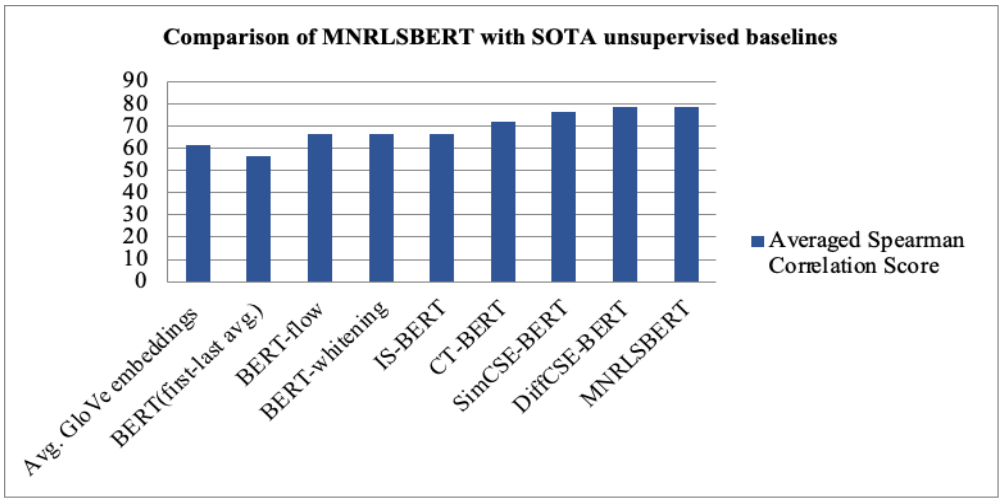
Tables 1 and 2 shows the evaluation results of our models compared with unsupervised and supervised baseline models respectively. It represents the performance of each model in terms of Spearman's rank correlation values (x100).

The results from Table 1 indicate that MNRLSBERT and MNRLSRoBERTa outperform the previous SOTA unsupervised baselines with a clear margin (except DiffCSE, where it shows nearly equivalent performance). It is worth mentioning here that DiffCSE is trained on 106 sentences randomly sampled from English Wikipedia with an average running time of 3-6 hours whereas our models are trained with comparatively less data with an average training time of less

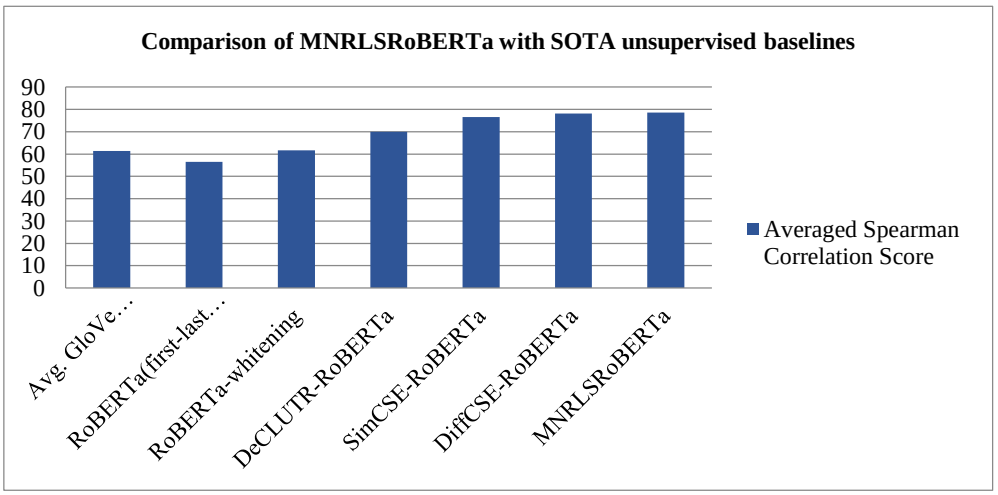
than 2 hours.

Our models show a substantial improvement over previous SOTA-supervised SEMs like InferSent-GloVe [8], Universal Sentence Encoder [9], and SBERT/SRoBERTa [5] as shown in Table 2. For BERT and RoBERTa as underlying transformers, the corresponding models SimCSE-BERT and SimCSE-RoBERTa are giving the highest performance. This is because supervised SimCSE is trained on NLI data under a contrastive learning framework with a greater number of epochs for longer. In contrast, our models are showing competitive results despite relatively less training time. Nevertheless, our models are the second-best performers among the supervised baselines.

Figures 6 and 7 graphically represents the averaged Spearman correlation score of MNRLSBERT/ MNRLSRoBERTa compared with SOTA unsupervised and supervised sentence representation baselines respectively.

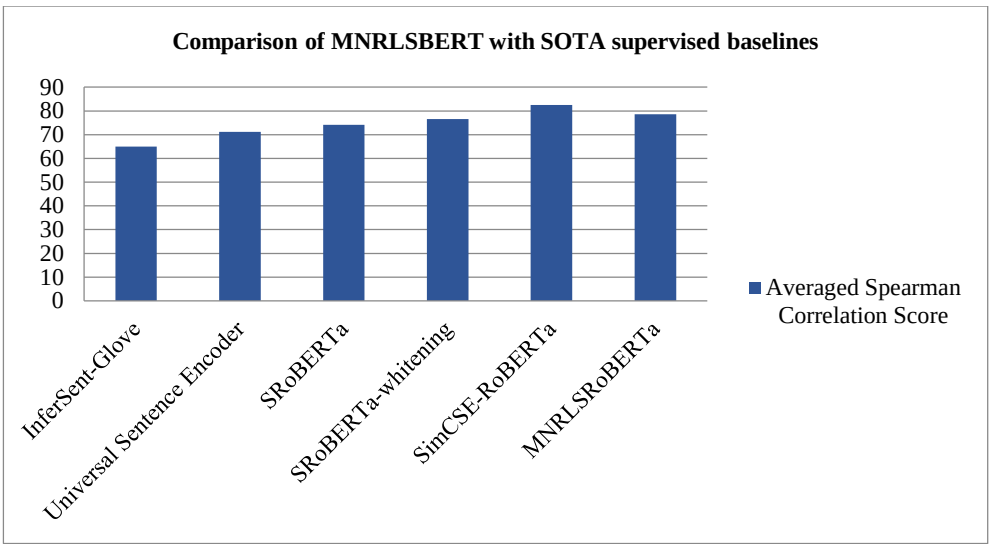


(a)

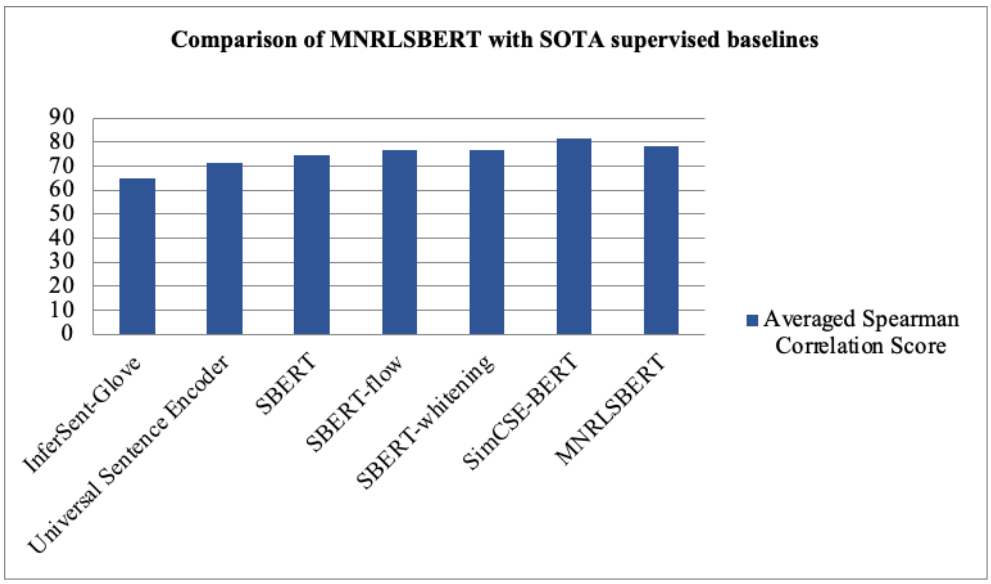


(b)

Figure 6. (a) Comparison of MNRLSBERT with SOTA unsupervised baselines, **(b)** Comparison of MNRLSRoBERTa with SOTA unsupervised baselines



(a)



(b)

Figure 7. (a) Comparison of MNRLSBERT with SOTA supervised baselines, **(b)** Comparison of MNRLSRoBERTa with SOTA supervised baselines

VI. Conclusion and Future Scope

The cross-encoder BERT can provide reliable results for sentence pair regression task but due to its high computing demands, it does not scale well for applications like semantic search and clustering. SBERT is the first transformer based pre-trained SEM that drastically reduces the computational inference time from many hours to a few seconds. It has fine-tuned BERT/RoBERTa on NLI data by applying Siamese network and used SL function as primary method to train the model. In this paper, we have contributed MNRLSBERT and MNRLSRoBERTa, as more efficient novel variants of SBERT and SRoBERTa respectively using MNRL function, a training objective that works on the principle of contrastive learning. Each model is based on original SBERT architecture and is significantly outperforming many previous SOTA baselines for sentence representation. The results of our research work show empirically that our models are best suited for similarity search applications.

In this study, we trained our models with anchor-positive pairs only, we have not utilized the hard negatives. Also, due to limited computational resources, we have selected the base models of BERT/RoBERTa as underlying transformers. We believe that their large versions are likely to give better performance [38-41]. In the future, we would like to extend our research work using triplets (anchor, positive, negative) by introducing hard negatives in our training setup and work towards further improving the performance with optimal hyperparameter tuning [41-44].

References

[1] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding". In *Proceedings of the*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186, 2019. doi: 10.18653/V1/N19-1423.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. "Attention is all you need". In *Advances in Neural Information Processing Systems*, pp. 6000–6010, 2017. doi: 10.5555/3295222.3295349.

[3] C. Wei, Y.-C. Wang, B. Wang, and C.-C. J. Kuo. "An Overview on Language Models: Recent Developments and Outlook". Mar. 2023, [Online]. Available: *http://arxiv.org/abs/2303.05759*.

[4] K. Taneja and J. Vashishtha. "Comparison of Transfer Learning and Traditional Machine Learning Approach for Text Classification". In *Proceedings of the 9th International Conference on Computing for Sustainable Global Development, INDIACom 2022*, pp. 195–200, 2022. doi: 10.23919/INDIACom54597.2022.9763279.

[5] N. Reimers and I. Gurevych. "Sentence-BERT: Sentence embeddings using siamese BERT-networks". In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019. doi: 10.18653/v1/d19-1410.

[6] J. Pennington, R. Socher, and C. D. Manning. "GloVe: Global Vectors for Word Representation". In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543, 2014. doi: 10.3115/V1/D14-1162.

[7] F. Schroff, D. Kalenichenko, and J. Philbin. "FaceNet: A unified embedding for face recognition and clustering". In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07, pp. 815–823, 2015. doi: 10.1109/CVPR.2015.7298682.

[8] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes. "Supervised learning of universal sentence representations from natural language inference data". In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 670–680, 2017. doi: 10.18653/v1/d17-1070.

[9] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. John, N. Constant, M. Guajardo-Cespedes, Yuan, C. Tar, Y. Sung, B. Strope, and R. Kurzwell. "Universal Sentence Encoder" 2018, Accessed: May 22, 2023. [Online]. Available: *http://arxiv.org/abs/1803.11175*.

[10] B. Li, H. Zhou, J. He, M. Wang, Y. Yang, and L. Li. "On the sentence embeddings from pre-trained language models". In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 9119–9130, 2020. doi: 10.18653/v1/2020.emnlp-main.733.

[11] J. Su, J. Cao, W. Liu, and Y. Ou. "Whitening Sentence Representations for Better Semantics and Faster Retrieval" 2021, Accessed: Jul. 15, 2023. [Online]. Available: http://arxiv.org/abs/2103.15316.

[12] Y. Zhang, R. He, Z. Liu, K. H. Lim, and L. Bing. "An unsupervised sentence embedding method by mutual information maximization". In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1601–1610, 2020. doi: 10.18653/v1/2020.emnlp-main.124.

[13] J. Giorgi, O. Nitski, B. Wang, and G. Bader. "DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations". In *59th Annual Meet of Association for Computational Linguistics and Proceedings of 11th International Joint Conference on Natural Language Processing*, pp. 879–895, 2021. doi: 10.18653/V1/2021.ACL-LONG.72.

[14] F. Carlsson, E. Gogoulou, E. Ylipää, A. C. Gyllensten, and M. Sahlgren. "SEMANTIC RE-TUNING WITH CONTRASTIVE TENSION". In *Proceedings of 9th International Conference on Learning Representations*, 2021.

[15] T. Gao, X. Yao, and D. Chen. "SimCSE: Simple Contrastive Learning of Sentence Embeddings". In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, 2021. doi: 10.18653/v1/2021.emnlp-main.552.

[16] Y. S. Chuang, R. Dangovski, H. Luo, Y. Zhang, S. Chang, M. Soljačić, S. Li, W. Yih, Y. Kim, and J. Glass. "DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings". In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4207–4218, 2022. doi: 10.18653/v1/2022.naacl-main.311.

[17] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. "A large annotated corpus for learning natural language inference". In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, 2015 doi: 10.18653/v1/d15-1075.

[18] A. Williams, N. Nangia, and S. R. Bowman. "A broad-coverage challenge corpus for sentence understanding through inference". In *Proceedings of the Conference of the North American Chapter of Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 1112–1122, 2018. doi: 10.18653/v1/n18-1101.

[19] R. Hadsell, S. Chopra, and Y. LeCun. "Dimensionality reduction by learning an invariant mapping". In *Proceedings of. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, pp. 1735–1742, 2006, doi: 10.1109/CVPR.2006.100.

[20] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre. "SemEval-2012 Task 6: A pilot on semantic textual similarity". In *Proceedings of First Joint Conference on Lexical and Computational Semantics*, vol. 2, pp. 385–393, 2012

[21] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo. "*SEM 2013 shared task: Semantic Textual Similarity". In *SEM 2013 - 2nd Joint Conference on Lexical and Computational Semantics, Proceedings of the Main Conference and the Shared Task: Semantic Textual SimilaritySEM 2013 - 2nd Joint Conference on Lexical and Computational Semantics, Proceedings of the Main Conf*, vol. 1, pp. 32–43, 2013. Accessed: May 15, 2023. [Online]. Available: https://aclanthology.org/S13-1004.

[22] E. Agirre, C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, R. Mihalcea, G. Rigau, and J. Wiebe. "SemEval-2014 Task 10: Multilingual Semantic Textual Similarity". In *Proceedings of 8th International Workshop on Semantic. Evaluation (SemEval 2014) - co-located with 25th International Conference on*

Computational Linguistics, pp. 81–91, 2014. doi: 10.3115/V1/S14-2010.

[23] E. Agirre, C. Banea, C. Cardie, D. Cer, M. Deb, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, R. Mihalcea, G. Rigau, L. Uria, and J. Wiebe. "SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability". In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 252–263, 2015. doi: 10.18653/V1/S15-2045.

[24] E. Agirre, C. Banea, D. M. Cer, M. T. Diab, A. Gonzalez-Agirre, R. Mihalcea, G. Rigau, and J. Wiebe. "SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation". In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT*, pp. 497–511, 2016, doi: 10.18653/V1/S16-1081.

[25] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation", In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*, pp. 1–14, 2017. doi: 10.18653/V1/S17-2001.

[26] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli "A SICK cure for the evaluation of compositional distributional semantic models". In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pp. 216–223, 2014.

[27] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. "Distributed representations of words and phrases and their compositionality". In *Advances in Neural Information Processing Sysems.*, pp. 1–9, 2013.

[28] R. Kiros et al. "Skip-thought vectors". In *Proceedings of the Conference on Advances in Neural Information Processing Systems*, pp. 3294–3302, 2015.

[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient Estimation of Word Representations in Vector Space" Accessed: May 25, 2023. [Online]. Available: http://ronan.collobert.com/senna/.

[30] F. Hill, K. Cho, and A. Korhonen. "Learning distributed representations of sentences from unlabelled data". In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1367–1377, 2016. doi: 10.18653/v1/n16-1162.

[31] Y. Yang, S. Yuan, D. Cer, and S. Kong. "Learning Semantic Textual Similarity from Conversations". In *Proceedings of the third Workshop on Representation Learning for Natural Language Processing*, pp. 164–174, 2018. doi: 10.18653/V1/W18-3022.

[32] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger. "On measuring social biases in sentence encoders". In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 622–628, 2019. doi: 10.18653/v1/n19-1063.

[33] Y. Qiao, C. Xiong, Z. Liu, and Z. Liu. "Understanding the Behaviors of BERT in Ranking". 2019, Accessed: May 16, 2023. [Online]. Available: *http://arxiv.org/abs/1904.07531*.

[34] S. Humeau, K. Shuster, M.-A. Lachaux, and J. Weston. "Real-time Inference in Multi-sentence Tasks with Deep Pretrained Transformers". *arXiv Preprint. arXiv1905.01969*, 2019.

[35] R. Dangovski, L. Jing, C. Loh, S. Han, A. Srivastava, B. Cheung, P. Agrawal, and M. Soljačić. "EQUIVARIANT CONTRASTIVE LEARNING", 2022, Accessed: Jul. 15, 2023. [Online]. Available: https://github.com/rdangovs/essl.

[36] N. Reimers and I. Gurevych "Making monolingual sentence embeddings multilingual using knowledge distillation". In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 4512–4525, 2020. doi: 10.18653/v1/2020.emnlp-main.365.

[37] W. Tan and P. Koehn. "Bitext Mining for Low-Resource Languages via Contrastive Learning" 2022, [Online]. Available: *http://arxiv.org/abs/2208.11194*.

[38] Anguluri Rajasekhar, Ravi Kumar Jatoth, Ajith Abraham, Design of intelligent PID/PID speed controller for chopper fed DC motor drive using opposition based artificial bee colony algorithm, Engineering Applications of Artificial Intelligence, 29: 13-32, 2014.

[39] Hesam Izakian, Behrouz Ladani, Kamran Zamanifar and Ajith Abraham, A Particle Swarm Optimization Approach for Grid Job Scheduling, Third International Conference on Information Systems, Technology and Management, Communications in Computer and Information Science, Springer Verlag, ISBN 978-3-642-00404-9, pp. 100-109, 2009.

[40] Musrrat Ali, Millie Pant and Ajith Abraham, Simplex differential Evolution, Acta Polytechnica Hungarica, 6(5):95-115, 2009.

[41] Amit K.Shukla, Manvendra Janmaijaya, Ajith Abraham, Pranab K. Muhuri, Engineering applications of artificial intelligence: A bibliometric analysis of 30 years (1988-2018), Engineering Applications of Artificial Intelligence, 85: 517-532, 2019.

[42] Ivan Zelinka, Vaclav Snasel and Ajith Abraham, Handbook of Optimization: From Classical to Modern Approach, Intelligent Systems Reference Series, ISBN 978-3-642-30503-0, Springer Verlag Germany, 1100 p, 2012.

[43] Zahra Pooranian, Mohammad Shojafar, Jemal Abawajy, Ajith Abraham, An efficient meta-heuristic algorithm for grid computing, Journal of Combinatorial Optimization, 30(3): 413-434, 2015.

[44] Prithwish Chakraborty, Swagatam Das, Gourab Ghosh Roy and Ajith Abraham, On Convergence of the Multi-objective Particle Swarm Optimizers, Information Sciences,181(8):1411-1425, 2011.

Author Biographies

.

Khushboo Taneja is born in India on 29th May 1988. She has obtained her B. Tech degree in computer science and engineering from Guru Jambheshwar University of Science and Technology, Hisar (India) in the year 2009. She has done Master of Engineering in computer science and engineering from Thapar University, Patiala (India) from 2010-2012. Currently, she is pursuing PhD. in Computer Science and Engineering in data mining from Guru Jambheshwar University of Science and Technology (India).

Jyoti Vashishtha obtained her Master's degree in Computer Applications from Himachal Pradesh University, Shimla, HP (India) in the year 1994. She has done PhD. in Computer Science and Engineering from Guru Jambheshwar University of Science and Technology, Hisar (India) in 2014. Currently, she is working as professor in Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology, Hisar (India). Her major field of research is machine learning and data mining.

Saroj Ratnoo earned her master's degree in computing science from Birkbeck College, University of London in 1993. She obtained her PhD. from Jawaharlal Nehru University, New Delhi in 2010. She is working as a Professor in Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology, Hisar (India). Her field of research includes nature-inspired algorithms, data mining and machine learning.