IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 January 2024, accepted 1 March 2024, date of publication 11 March 2024, date of current version 25 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3375754

RESEARCH ARTICLE

Neural Network Meaningful Learning Theory and Its Application for Deep Text Clustering

ELNAZ ZAFARANI-MOATTAR[1], MOHAMMAD REZA KANGAVARI[2], AND AMIR MASOUD RAHMANI[1]

[1]Department of Computer Engineering, Islamic Azad University, Science and Research Branch, Tehran 1477893855, Iran
[2]Department of Computer Engineering, Iran University of Science and Technology, Tehran 13114-16846, Iran

Corresponding author: Mohammad Reza Kangavari (kangavari@iust.ac.ir)

ABSTRACT In this paper, a new theory to train neural networks is presented which is called ''Neural Network Meaningful Learning'' (NNMeL) theory. According to this theory, meaningful learning in an artificial neural network occurs when concepts are learned separately but get related to each other. This theory is theoretically supported by ''Azobel's cognitive theory of learning''. According to Azobel's theory, the most important effective factor in learning is previous learning, and meaningful learning occurs when a person consciously relates new knowledge to what they already knew. Also, a new model named ''Deep Clustering based on Sentence Similarity'' (DCSS) is proposed for topic detection. This model proposes to use similarity of sentences to produce sentence representation instead of using autoencoder and training it based on denoising. Also, a trainable framework based on NNMeL is presented. Many experiments conducted for evaluation. First of all, an experiment is arranged to check the correctness of the NNMeL theory. The results confirm that training ANN according to this theory (training concepts one after another), gets better results than training all concepts together. In the following experiments, the DCSS model, which is based on sentence similarity, is compared with another autoencoder based method not only by evaluation metrics but also by human evaluators. The results indicate a 6.6% improvement in accuracy and better human-evaluators satisfaction. Finally, the proposed model is compared with ten other methods in topic detection application. The results show that the proposed model is superior to the rest of the methods.

INDEX TERMS Neural network meaningful learning theory, topic detection, sentence similarity, deep text clustering.

I. INTRODUCTION

Topic detection is one of the most important issues in natural language processing. The exponential growth in data production due to the rapid expansion of social networks has made automatic topic detection increasingly crucial. Text clustering is one of the most important and effective ways to identify the topic. Because it enables us to have a better analysis of the high volume of posts by grouping them in small groups, i.e. clusters. Text clustering can be divided into two parts, short text clustering and long text clustering, according to the volume of text data. Short texts may contain one or more lines, such as the textual content of social networks and news headlines. Nowadays, a high proportion of the texts used in daily life are short texts because communication formats in social networks focus more on texts with one or a few lines. Clustering short texts is much more challenging than long texts because text clustering requires measuring the similarity between texts, while many similarity measurement methods for long texts are not effective for short texts. Because short texts cannot fully explain concepts like long texts. Also, a few words are used in short texts, so we cannot use methods such as term frequency.

Deep clustering is a completely new solution that has recently been proposed in the field of clustering. While deep learning was introduced several years ago in AI research,

The associate editor coordinating the review of this manuscript and approving it for publication was Alicia Fornés.

deep clustering has only recently gained attention due to the challenges in applying deep learning to clustering tasks. Initially, deep clustering primarily focused on numerical data and later expanded to include image data. More recently, deep clustering has been applied to text data. According to [1], researches on text clustering are focused on three main areas: Deep learning-based feature extraction methods, BiLSTM-based ELMO model and clustering algorithms based on autoencoder.

Most deep clustering methods are based on an autoencoder neural network, which is used to map input vectors to latent space vectors. It can be said that these latent space vectors are, somehow, features extracted from the input vectors. Most existing methods directly use these feature vectors for clustering, but there are other papers that apply another mapping to them before clustering. How to train autoencoder is a challenge in this field. Almost all existing methods, give the sample xi as desired output and the noisy version of sample xi as input to the autoencoder in the training phase. In this way, the autoencoder network learns how to denoise and produces the latent space vector during this operation. But in our opinion, the production of latent space vectors during denoising is not close to how human does. Therefore, in this article, we will present a new method based on sentence similarity for training the network and generating latent space vectors. In our opinion, humans use the similarity of sentences for clustering rather than generating concepts based on denoising.

Our thesis in this paper is that during the language learning process, from childhood to adulthood, humans first learn the conceptual similarity of words and sentences, then clustering. In other words, first the learning of "sentence similarity" occurs and then the learning of "clustering" takes place based on the learned concept of "sentence similarity". The theoretical support of our theory is "Azobel's cognitive theory of learning". According to Azobel's theory, the most important effective factor in learning is previous learning, and meaningful learning occurs when a person consciously relates new knowledge to what he/she already knew [1]. Therefore, in this article, we have designed a framework in which the proposed model first learns the concept of "sentence similarity" and then, based on this concept, it learns "how to separate vectors for clustering"; i.e., it relates the new concept to the previously learned concept. We call this theory as "Neural Network Meaningful Learning" (NNMeL) theory.

As a summary, this paper proposes a new model based on sentence similarity and deep clustering for topic detection. This model is based on "Azobel's cognitive theory of learning". Also, a model is presented for the first time, to make deep neural network trainable for generating sentence representation vectors based on sentence similarity. The general idea of the proposed model is to map the input sentences to a contextual latent space at the first step. The similarity of the sentences is used for this manner. Then, another neural network is used to create another map from contextual space to another space in which the vector of input sentences are more separable, so that clustering can be done better. This idea is novel both in theory and in modeling. Theoretically, the novelty of this approach lies in the construction of the representation vector of the posts in a way that more attention is paid to the keywords of the sentence, the sequence of words is included in the representation vector of the sentence, and importantly, the representation vector considers both similarity of the sentences and separability of vectors. In terms of modeling, its important novelty is to provide a trainable model, i.e., one that automatically generates the representation vector with the aforementioned features.

The main contributions of this article are:

- Providing a theory named "Neural Network Meaningful Learning theory" which is based on "Azobel's cognitive theory of learning" to enhance training of deep neural networks. This theory is a milestone in machine learning because this leads us to train neural networks for more complicated tasks than what we have currently.
- Presenting a new model based on "sentence similarity" to generate sentence representation vectors instead of using "autoencoder". This is the first time that sentence similarity is used to generate sentence representation vectors and a trainable model is presented for it.
- Providing a framework to make trainable neural network for "generating latent vectors based on sentence similarity" rather than "generating latent vectors based on denoising". Most of previous works generate latent space vectors based on denoising but this paper generates them based on sentence similarity which is very similar to what human do and is compatible with clustering task.
- Separating roles (versus mixing all together in one vector). In the presented model, different vectors are used for different roles i.e. "sentence representation vectors", "latent space vectors" and "vectors used for clustering" are separated from each other. Our experiments confirm that better results are obtained by using separate vectors for each role. This is an important result because it shows that in order to achieve more complex roles, it is better to break them into several simpler roles and train each one separately, then combine them into a more complex network. This result paves the way for constructing more complex networks to handle more complicated tasks.

II. RESEARCH BACKGROUND

As mentioned before, "Azobel's cognitive theory of learning" is the theoretical basis of the model presented in this article. In 1963 the first attempt was made in [2] to present a cognitive theory of *meaningful* verbal learning as opposed to *rote* verbal learning [3]. However, this theory is still a popular theory in psychology and some attempts are done to update it such as [4]. David Azobel is a psychologist who distinguishes between rote learning and meaningful learning. According to Azobel's theory, the most important effective factor in learning is previous learning. Meaningful learning occurs when a person consciously relates new knowledge to what he/she already knew. When meaningful learning occurs, some changes are made in entire cognitive structure of the human mind, which change both the existing concepts and the existing relationships between them. This is why meaningful learning possesses greater memorization and generalization capabilities compared to non-meaningful or rote learning. In non-meaningful learning, the information is simply stored in the memory without any connection with the cognitive structure. There are many research evidences that indicate the superiority of using concept maps over the common methods of teaching and learning. (For example: [5], [6], [7].) According to Azobel's theory, learning is meaningful when new subjects are related to previously learned subjects and expand or change them. Therefore, in the proposed model of this article, first ''sentence similarity'' is learned, then ''clustering'' is learned by relating it to ''sentence similarity''.

There are two different views in text processing: Classification [8] vs. Clustering. Also, in the field of NLP, RNN and LSTM neural networks were raised respectively [9].

''Deep text clustering'' is a new category in topic detection. Various categorizations were previously presented in the literature for topic detection methods [10], [11], [12], [13]. One of the most prominent categories in these categorizations was ''clustering''. In 2021 [14], a new category called ''methods based on deep learning'' was added to the previous categorizations. Combination of this category with ''Clustering'' category in the field of text, created a new category named ''Deep Clustering of Texts''. In recent years, the concept of transfer learning in the field of text and consequently the use of pre-trained language models and embedding methods for text representation have been of great interest. Since the input text cannot be directly used in clustering algorithms, word embedding methods were used along with clustering algorithms. Three components of embedding methods, distance metrics and clustering algorithms influence each other in clustering. In [15], the effect of these three components has been investigated simultaneously and the efficiency of five new embedding methods, two distance metrics and five clustering algorithms have been considered. In [16], fourteen methods of text representation in Twitter data have been empirically investigated, ten of which are unsupervised. In [17], a framework called ''STC^2'' is presented for the first time for short text clustering, in which a self-learning convolutional neural network is used.

The efficiency of the clustering algorithm is highly dependent on the quality of the data representation. Deep clustering methods try to map data by deep neural networks to a space in which the data can be easily separated by clustering algorithm. This type of data representation is called clustering-friendly representation [18]. Another method used for deep text clustering is the autoencoder neural network [19], [20], which is one of the most prominent unsupervised representation learning algorithms [21]. This neural network creates a non-linear mapping from the data space to a latent space in order to reduce the dimensions. The latent space should have smaller dimensions than the data space while containing the most salient features of the data [22]. Most of up to date deep text clustering methods are based on autoencoder neural network which is used to map input vectors to latent space vectors. Many existing methods directly use these feature vectors for clustering (e.g. [23], [24]) and a limited number apply another mapping to them before clustering (e.g. [25]). Most autoencoder networks provide the sample x_i as the desired output and the noisy version of the sample x_i as the input to the autoencoder during the training phase due to the lack of labeled data. By doing so, the autoencoder network learns how to denoise and produces the latent space vector during this operation. In [26], convolutional autoencoder is used for deep representation learning in clustering of health tweets. The only existing work on Persian language is [27] which is done on ''Barez'' dataset. This dataset includes the suggestions that have been given to improve the quality of the ''Barez'' company products. The ParsBert embedder is used for word embedding, stacked autoencoder for feature extraction, and k-means algorithm for clustering.

III. PROPOSED METHOD

In this paper, a new method based on ''Azobel's cognitive theory of learning'' named ''Deep Clustering based on Sentence Similarity'' (DCSS) is proposed for topic detection. The main idea of this method is inspired by the clustering process in humans. For clustering, humans read individual posts and, based on the similarity of the sentences, relate them to one of the previous posts or a group of previous posts, and thereby, divide the posts into categories. Therefore, we present the idea of using sentence similarity to generate latent space vectors in this article for the first time. That is, we obtain the latent space vectors using sentence similarity rather than denoising the input. Another novelty of this paper is to provide a framework to make these vectors trainable, because these vectors cannot be trained directly.

The model of the proposed method is shown in Figure 1. This model consists of four components: Contextualized Sentence Encoder (CSE), Sentence Similarity Classifier (SSC), Vector Separator Network (VSN) and Clustering Algorithm (CA). In the remaining of this section, we will first explain the details of each of these components and how to train them and then, the novelty of the proposed model.

CSE's duty is to provide a Contextualized Sentence Representation (CSR). This representation should not only represent the text and the context of the sentences and words it contains, but also should be in a manner that similar sentences have close representation and sentences that are thematically far from each other, obtain distant representation. Therefore, it can be said that this component is a mapping from the space of words and sentences to a contextual space which is, similar to the embedding of words, indicating the meaning of sentences and the similarity of sentences. In the CSE component, the attention mechanism has been used to give more focus and emphasis on important words and give less weight to other

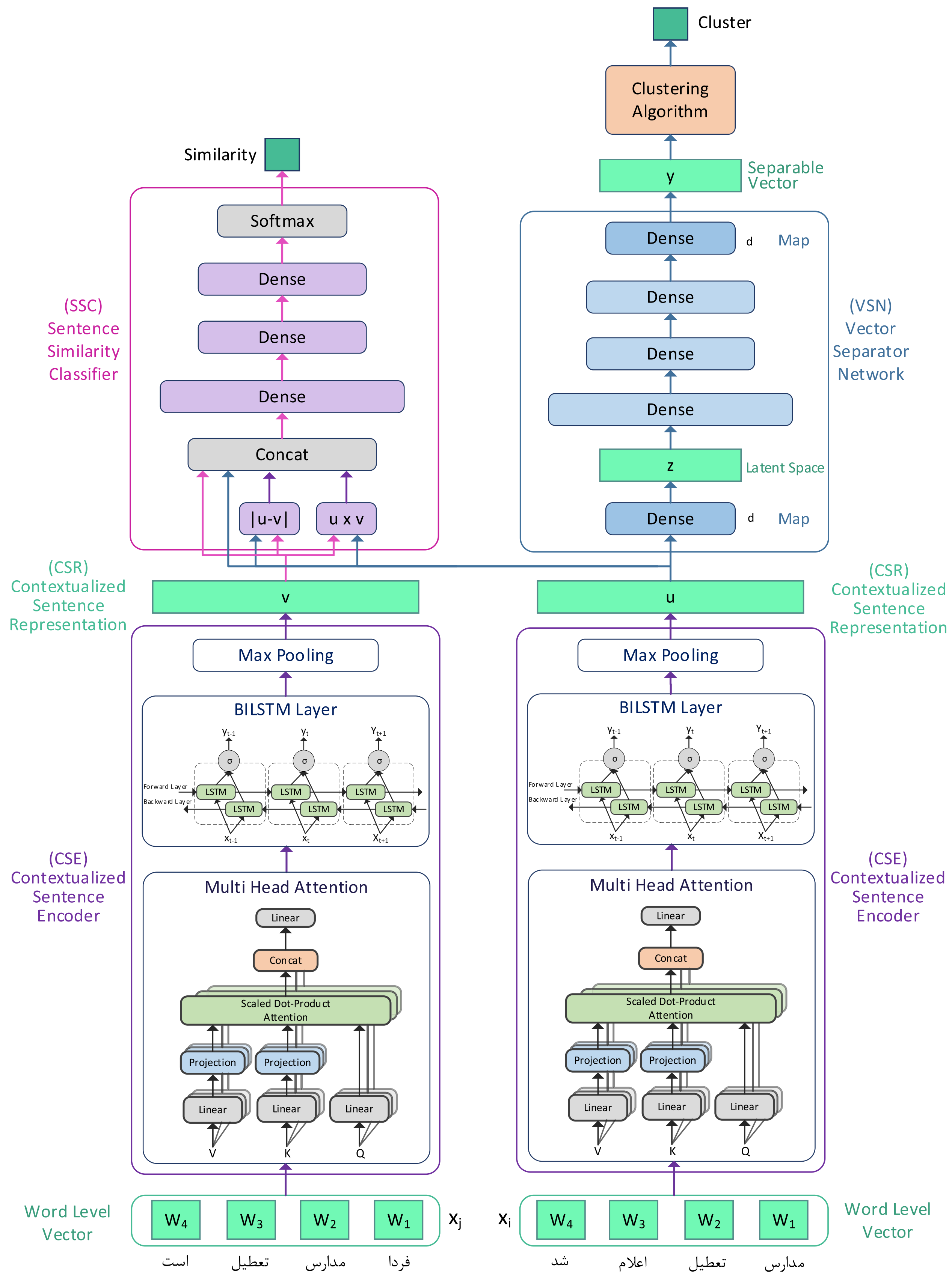

FIGURE 1. **Proposed model. DCSS: Deep clustering based on sentence similarity.**

words. Also, LSTM is used to preserve the order of words and to take into account the context in which the words are located. Therefore, it can provide a contextualized sentence embedding. Since the contextualized space resulting from CSE is a latent space, it cannot be trained directly. Therefore, in this article, we have presented the idea of using the Sentence Similarity Classifier (SSC) component. According to this idea, a pair of sentences are passed through the CSE component and then are entered into the SSC component to determine their similarity or dissimilarity. The combination of these two components in such way makes it trainable, because now it is possible to train these two components together using a dataset that determines the similarity of the sentences. Training these two components together results in the creation of a contextualized latent space. Therefore, the first duty of SSC is to make CSE trainable. The second duty of the SSC component is to consider the similarity in the generated u vectors. In other words, the attention of the SSC component to the similarity of the sentences causes similar sentences to have similar u vectors and dissimilar sentences to have distant u vectors. Therefore, it can be said that the SSC component has two duties: first, to make the CSE component trainable, and second, to establish similarity between the u vectors produced by the CSE component. It is worth mentioning that the second duty is the most important duty of this component and the proposed novelty of the present article.

The third component is the vector separator network. The duty of this component is to make the representation vectors of the sentences separable. As mentioned above, the u vectors resulting from the CSE component provide a representation of the sentences which represents the similarity of the sentences due to the effect of the SSC component. But these vectors are not necessarily separable. The duty of VSN is to create a mapping that converts u vectors into separable vectors y, enabling the samples to be easily and effectively separated by the clustering algorithm.

Finally, the fourth component is the clustering algorithm component. The duty of this component is to cluster posts. For this purpose, any of the available clustering algorithms can be used.

One of the novelty of the proposed model lies in the idea and theory of how to obtain the representation vectors of the sentences. On the one hand, this idea focuses on the construction of sentence vectors from words, while on the other hand, it focuses on the separability of these vectors based on the similarity of the sentences to improve the accuracy of clustering. The usual method in constructing the vector of sentences from words is the averaging method. Our experience in implementing different methods ([15]) showed that this method is not highly efficient. This point is worth mentioning to justify the reason for this result which is the effective words in determining the topic are usually in the minority compared to other words in the post. Since averaging tends toward the majority, the resulting sentence vectors are closer to the ineffective majority than to the effective minority. This causes the post vectors to be close to each other, making them difficult to cluster. Therefore, a neural network is used in this article in which both the order of the words and attention to the more important words are considered. The attention mechanism has been used in the CSE component for this purpose. Using the attention mechanism allows the network to assign more weight to more important words. On the other hand, the proposed model constructs sentence vectors based on the similarity of sentences, which is not observed in other papers and is one of the main novelties of this article. (The way the components are combined and the way the network is trained accomplish this.) Previous works employ an encoder to map the input vector (x) to the latent space (z). In order to make it trainable, they use a decoder to reverse map from the latent space to the space of input vectors (z space to x space) so that an autoencoder is formed and can be trained. The pair $<x', x>$ is used as training sample, where x' is the noisy version of x. That is, the noisy vector is given to the input of the autoencoder and the original vector is requested as the output of the autoencoder. This training method makes the network resistant to possible noises in its input. Oppositely, proposed model employs two inputs instead of single input, and the similarity of the input pair is used to train it. This causes u-vectors to be close to each other for similar inputs and far-apart for dissimilar inputs, resulting is separable vectors based on similarity. This, in turn, helps to cluster the input posts more effectively. Therefore, the main idea of this paper is to make the model trainable and enable it to generate sentence representations based on their similarity. In the ''experiments and results'' section, we will find that this idea plays an effective role in increasing the accuracy of the topic detection (clustering of posts).

IV. DATASET

The dataset used in this research is the Sep_TD_Tel01 dataset [28] in Persian. This dataset can be used for topic detection. This dataset is collected from the Telegram social network without any special restrictions such as keywords, and therefore it truly has a data stream nature which is the most important feature of this dataset. The dataset is not biased because all the messages posted in public channels and groups are collected. When messages are collected using keywords, the dataset is biased towards those keywords. We have used this dataset not only as a golden standard for evaluating topic detection, but also for preparing sentence similarity dataset. The similarity dataset is prepared as follows: any pair of posts in the dataset are taken. The similarity of these pair of posts is considered as one if they are from the same topic and their similarity is considered as zero if they come from different topics.

V. TRAINING PROCESS

In the training process, three components must be trained, and to train each component, a specific loss function must be defined. Let X, U, Z, S, and Y denote the set of input vectors, the set of sentence representation vectors, the set of latent space vectors, the set of similarities, and the set of output vectors, respectively. V is similar to U. Let $x_1, x_2, \ldots, x_n$ are the set of training samples. The objective is to obtain a loss function for each of the components which enable their training and lead them to the goal which is considered for each component.

The SSC component can be considered as a map from the sentence representation space to the sentence similarity space, denoted as $g : U \times U \to S$. For the SSC component, mean squared error (MSE) between the actual similarity of the sentences and the similarity obtained by the SSC component is used. Having $s_i \in S$ as the desired similarity between the pair of input sentences, the loss function of the SSC component is defined as follows:

$$L_{SSC} = \frac{1}{n} \sum\nolimits_{i=1}^{n} (g(u_i, v_i) - s_i)^2 \tag{1}$$

The VSN component can be considered as a map from the sentence representation space to the separable vectors space as $h : U \to Y$. Usually, criteria such as silhouette are calculated on the output of the clustering algorithm to measure the quality of clustering. The silhouette score is better when the distance between within-cluster samples is small and the distance between the clusters is large. But in this case, the loss function cannot be defined on the output of the clustering algorithm. The reason is if we want to train the network by a loss function defined on the output of the clustering algorithm, the derivation of the loss function with respect to the clustering algorithm must be calculated to use it to train the previous layers, which is impossible. Therefore, the output of the VSN component should be the basis for error calculation. Therefore, we use the following loss function to train the VSN component:

$$L_{VSN} = H(P) + KL(P||Q) \tag{2}$$

where $H(P)$ is the entropy function and $KL(P||Q)$ is the Kullback–Leibler divergence function, which are defined as follows:

$$H(P) = -\sum\nolimits_i \sum\nolimits_j p_{ij} \log(p_{ij}) \tag{3}$$

$$KL(P|Q = -\sum\nolimits_i \sum\nolimits_j p_{ij} \log(\frac{q_{ij}}{p_{ij}}) \tag{4}$$

where P and Q are probability distributions. In order to extract the data structure, the distance between data pairs in both the latent space and the output vector space are converted into probabilities. The t-Student distribution is used for this purpose as follows:

$$p_{ij} = \frac{\left(1 + \left\|\hat{z}_i - \hat{z}_j\right\|^2 / \alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{k \neq l} \left(1 + \left\|\hat{z}_k - \hat{z}_l\right\|^2 / \alpha\right)^{-\frac{\alpha+1}{2}}} \tag{5}$$

$$q_{ij} = \frac{\left(1 + \left\|\hat{y}_i - \hat{y}_j\right\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \left\|\hat{y}_i - \hat{y}_j\right\|^2\right)^{-1}} \tag{6}$$

where $\hat{y}$ and $\hat{z}$ are the vectors obtained for y and z by the proposed model (Figure 1) and α is the degree of freedom of the t-Student distribution. $\alpha = 2d$ is considered in this research, where d is the dimension of the latent space. The use of the loss function defined in (2) has two effects on the model: 1) Since P varies during the training process, the use of entropy makes the network to generate latent space vectors in a way that the dispersion of the generated vectors decreases in order to reduce the entropy of P. Because reducing the dispersion of generated vectors leads to reducing the entropy of P. If the p_{ij} probabilities are approximately equal, the entropy increases. Entropy decreases when some p_{ij} probabilities are low and others are high. It means that there is a big difference between the p_{ij} probabilities. Since p_{ij} are obtained from distance of vectors $\hat{z}_i$ and $\hat{z}_j$ in Z space, entropy decreases when the distance of some pairs of vectors is small and the distance of others is high. This means that in order to decrease the entropy, the intra-cluster distance should be reduced and the inter-cluster distance should be increased. 2) Using the Kullback–Leibler divergence function as a part of the loss function and its reduction causes the divergence of Q distribution from P to be reduced to a minimum, i.e., to preserve the statistical characteristics of the latent space vectors (Z) in the output (Y). This avoids the distortion of the latent feature space during the h mapping process. Therefore, using the L_{VSN} loss function as defined in (2) makes both the main goals of clustering, i.e. reducing the intra-cluster distance and increasing the inter-cluster distance, to be achieved. This is the objective that the use of the silhouette loss function in regular clustering leads us to.

Finally, we come to how to train the CSE component. This component can be viewed as a mapping from the input words' vector space to the sentence representation space, which can be expressed as $f : X \to U$. Since the CSE is trained through two components, VSN and SSC, its loss function is considered as the sum of the loss functions of these two paths. i.e.:

$$L_{CSE} = L_{SSC} + L_{VSN} \tag{7}$$

Now that the loss function for each component is determined, it is time to discuss how to train the network. As mentioned, the DCSS model uses the SSC component to train the CSE component in order to obtain sentence representation vectors that represent their similarity. After that, the VSN component can map these vectors into a separable space by the clustering algorithm. Therefore, DCSS training is also done in two phases. In the first phase, only the CSE and SSC components are trained, and in the second phase, the entire network. But during application, only CSE and VSN components are used. The process is: The words of each input sentence are given to CSE to generate the corresponding u vector; then this vector is entered into VSN and the resulting y vector is used for clustering.

In the first phase of training, the CSE and SSC components are trained by the similarity dataset described in the previous section. The ith training sample in the similarity dataset includes two sentences Snt_i^1 and Snt_i^2 as input and their similarity (Sim_i) as output. These two sentences are mapped

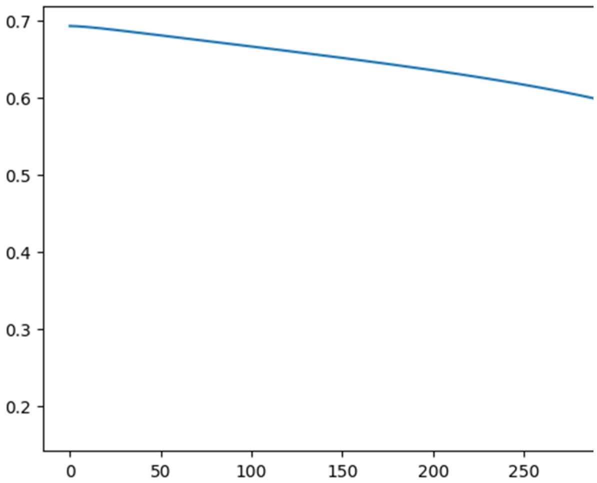

FIGURE 2. Improvement of the model during training in phase 1 based on loss function defined for phase 1. (In the same scale of Figure 3, to compare.)

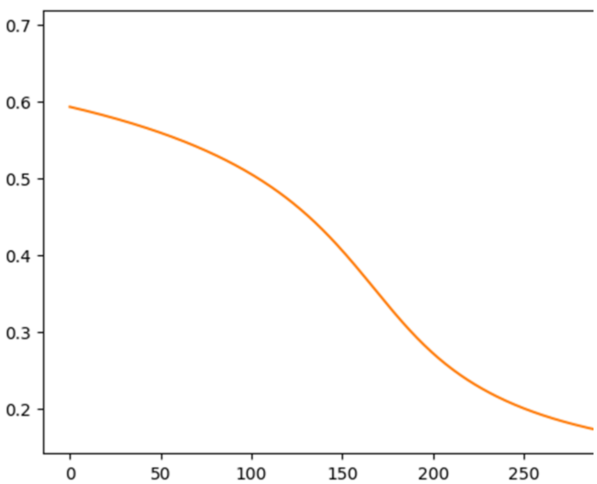

FIGURE 3. Improvement of the model during training in phase 2 based on loss function defined for phase 2. (In the same scale of Figure 2, to compare.)

by the CSE component to their corresponding representation vectors i.e. u_i and v_i. These two vectors are then fed into the SSC component to determine their similarity or dissimilarity. The value of Sim_i in the training sample is considered as the expected value from the SSC network to calculate network error, and CSE and SSC networks are trained together.

In the second phase of training, all three components of CSE, SSC and VSN are trained by Sep_TD_Tel01 dataset. At this stage, the weights obtained in phase 1 for CSE and SSC networks are considered as initial value. Therefore, phase 2 training is the main training for VSN network, but it is considered as fine tune for CSE and SSC networks.

Three strategies for training the network in the second phase can be envisioned. In the first strategy, both CSE and SSC networks are trained together with VSN, i.e. both CSE and SSC are involved in the fine tuning process. In the second strategy, fine tune is performed on CSE, but SSC is not included in the training process and maintains the same initial weights; i.e. it keeps its pretrained weights. In other words, in this strategy, the SSC component remains the same as phase 1 and does not receive new training. In the third strategy, both CSE and SSC networks are used as pretrained and fine tune is not performed on them. Therefore, in this case, only the VSN network will be trained and the rest of the model will keep the training results of phase 1. Obviously, in all three strategies, the VSN network will be trained in phase 2, because the training of this network is only done in phase 2. It should be noted that a fourth strategy is also envisioned here, where CSE maintains its initial weights (pretrained), but SSC is involved in training (fine-tuning). But this strategy is practically useless, because VSN is independent from SSC, and therefore, changing the weights of SSC and consequently changing the output of SSC will not affect the performance of VSN. Therefore, this strategy is abandoned.

TABLE 1. Hyperparameter settings for DCSS.

Parameter	Value
Latent Space Size	10
Dense 1 units VSN	2000
Dense 2 units VSN	500
Dense 3 units VSN	500
Separable Vector Size	10
Dense 1 units SSC	1000
Dense 2 units SSC	200
Dense 3 units SSC	2

VI. EXPERIMENTS AND RESULTS

The theory proposed in this article will be evaluated in this section. First experiment checks correctness of the theory. The proposed model is compared with other methods in the second and fourth experiments from clustering and topic detection point of view. The proposed model is evaluate by human in the third experiment. The hyperparameter settings used in the experiments for the DCSS model are provided in Table 1.

Experiment 1: The objective of first experiment is to ascertain the correctness of the idea presented in this article. As previously mentioned, the main idea of this paper is to generate a representation of sentences based on their similarity and to make it trainable, and we have done this based on "Neural Network Meaningful Learning" theory. Now we want to answer the question whether this idea really leads to the improvement of clustering in practice? In other words, this experiment simultaneously measures the correctness of "meaningful neural network learning" theory and the correctness of the idea of representation generation based on similarity. For this purpose, we trained the model with the three strategies mentioned in the previous section. Figure 2 and Figure 3 show improvement of the model during the training of the model in phase 1 and 2 respectively based on loss functions define for each phase. After training the model, the evaluation results of the DCSS method in all three

TABLE 2. **Examining the correctness of "Neural network meaningful learning" theory and, simultaneously, examining the impact of the idea of "training the model through sentence similarity" on the quality of topic detection. "Pretrained" means the component (CSE or SSE) preserves the weights of phase 1 of training, and "Fine tune" means that the component undergoes training in phase 2 as well (according to the three strategies for model training in phase 2). The results indicate the evaluation of the output of the DCSS method (topic detection quality) using accuracy and NMI criteria. Adam is used as optimizer.**

CSE (Contextualized Sentence Encoder)	SSC (Sentence Similarity Classifier)	Optimizer	Accuracy	NMI
✓ Fine tune ☐ Pretrained	✓ Fine tune ☐ Pretrained	✓ Adam ☐ SGD-MOM	0.511	0.756
✓ Fine tune ☐ Pretrained	☐ Fine tune ✓ Pretrained	✓ Adam ☐ SGD-MOM	0.496	0.754
☐ Fine tune ✓ Pretrained	☐ Fine tune ✓ Pretrained	✓ Adam ☐ SGD-MOM	**0.562**	**0.782**

TABLE 3. **Examining the idea of the article by topic detection. This table is similar to Table 2, except that SGD-MO Optimizer is used.**

CSE (Contextualized Sentence Encoder)	SSC (Sentence Similarity Classifier)	Optimizer	Accuracy	NMI
✓ Fine tune ☐ Pretrained	✓ Fine tune ☐ Pretrained	☐ Adam ✓ SGD-MOM	0.642	0.825
✓ Fine tune ☐ Pretrained	☐ Fine tune ✓ Pretrained	☐ Adam ✓ SGD-MOM	0.664	0.827
☐ Fine tune ✓ Pretrained	☐ Fine tune ✓ Pretrained	☐ Adam ✓ SGD-MOM	**0.679**	**0.846**

TABLE 4. **Evaluation of the proposed method from clustering point of view: Comparing the results of the proposed method (DCSS) with the baseline deep clustering method (DAEC) on the Sep_TD_Tel01 dataset by Accuracy and NMI metrics.**

Method	Accuracy	NMI
DAEC [24] (Deep clustering method 2022)	0.613	0.783
DCSS (Proposed method - best configuration)	0.679	0.846

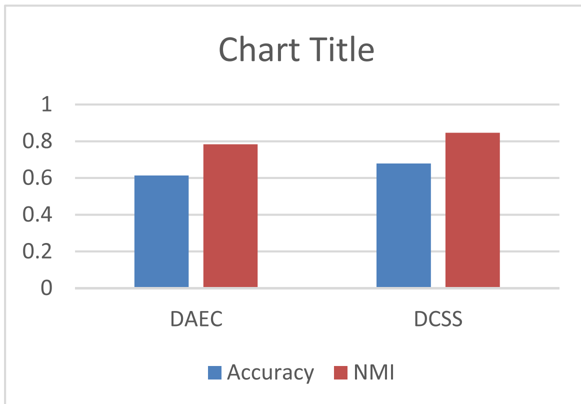

FIGURE 4. **Comparing the results of the proposed method (DCSS) with the baseline deep clustering method (DAEC) by Accuracy and NMI metrics.**

strategies by accuracy and NMI criteria using the golden standard dataset are shown in Table 2 and Table 3. It is clear that the third strategy achieves the best results in both tables. In other words, the vector u obtained in phase 1 (i.e. the vector resulting from the similarity of the sentences) gets the best results, and any change in the CSE weights in phase 2 (which leads to a change in the vector u) reduces the efficiency. In other words, the best results are achieved when the u vectors are obtained based on similarity, and any change in them in the later phases leads to worse results. Therefore, the hypothesis of this experiment is confirmed. That is, the idea of generating a vector of sentences through the similarity of sentences is a correct and effective idea. In addition, the best results are achieved when the learning of similarity is consolidated prior to learning separation and remains unchanged.

Experiment 2: After confirming the efficiency of the main idea of this research, it is time to measure the efficiency of the whole model in comparison with other methods. For this purpose, the proposed DCSS method was compared with the DAEC method [24]. The DAEC method is a completely new and up-to-date deep clustering method. For this purpose, the DEAC method was implemented and tested with the same Persian dataset (Sep_TD_Tel01). The comparison results are presented in Table 4 and Figure 4. According to this table, the proposed DCSS method has improved the accuracy by 6.6% and NMI by 6.3% compared to the DAEC method.

Experiment 3: Although the second experiment shows the efficiency of the proposed model by accuracy and NMI criteria, however, to ensure the efficiency of the proposed model, we decided to evaluate the output of the proposed model by human. Four human evaluators familiar with the topic detection process were employed for this purpose. First, the set of posts of each window from Sep_TD_Tel01 dataset was provided to the evaluators to study. Then, the output of one of these two methods (DAEC method and the proposed method) was selected randomly and was provided to them and they

TABLE 5. Comparison of human evaluation results for the proposed method (DCSS) and the deep clustering baseline method (DAEC) on the Sep_TD_Tel01 dataset. Each window was evaluated qualitatively by four evaluators on a scale of 1 (poor) to 5 (excellent). The score of each method is the average of scores given by all evaluators across all windows. The normalized score is obtained by dividing the score by 5, with a maximum value of 1.

Method	Score	Normalized Score
DAEC [24] (Deep clustering method 2022)	2.896	0.579
DCSS (Proposed method - best configuration)	3.487	0.697

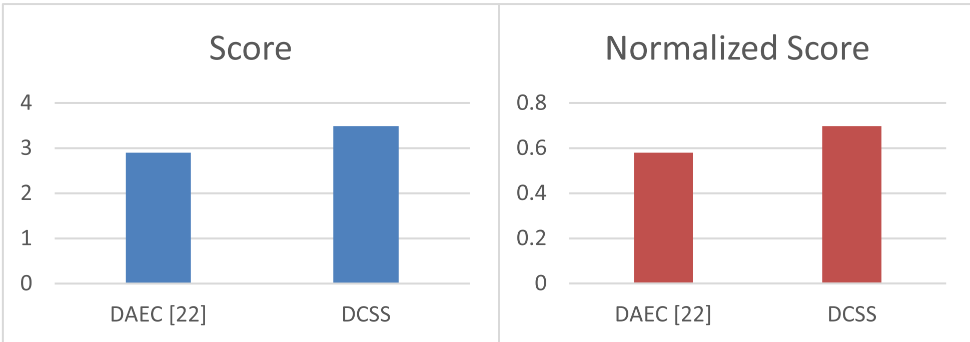

FIGURE 5. Comparison of human evaluation results for the proposed method (DCSS) and the deep clustering baseline method (DAEC).

were asked to evaluate and score the quality of the output. Afterwards, the output of the other method was given to them and they were asked to evaluate and rate its quality. The order of output delivery for each evaluator was randomly determined to prevent bias towards one of the methods and to avoid any mental pre-judgment among the evaluators, ensuring fair results. The evaluation was conducted by assigning scores to the quality of the output of each method in each window. The assessment was conducted qualitatively, within a range of 1 (indicating poor) to 5 (indicating excellent). Averaging the score of all evaluators over all windows determines the score of each method. The results obtained from this experiment are shown in Table 5 and Figure 5. We have also calculated the normalized score to have a better understanding of the scores. The normalized score is obtained by dividing the score by 5, with a maximum value of 1. According to this table, the proposed model has also been able to achieve a higher score in terms of human evaluation.

Experiment 4: The purpose of the fourth experiment is similar to the second experiment, to measure the efficiency of the whole model compared to other methods. The primary distinction between this experiment and the second one lies in the case study of these two experiments. In the second experiment, the proposed method was compared with a clustering method, i.e. the comparison was made from the perspective of clustering. However, in this experiment, the proposed method is compared with topic detection methods, i.e. the comparison is done from the perspective of topic detection. The second distinction lies in the utilization of new criterion FS in this experiment. The FS criterion focuses on posts in a particular topic; hence, the quality of topic detection is assessed based on its relevant posts. In this experiment, the proposed DCSS method is compared with ten topic detection methods [29]. The results are shown in Table 6. According to the results, the proposed DCSS method also performs better than other topic detection methods. Figure 6 demonstrate these results too.

TABLE 6. Evaluation of the proposed method from topic detection point of view: Comparing the results of the proposed method (DCSS) with ten topic detection methods on the Sep_TD_Tel01 dataset by ClassFS, ClusterFS, and MeanFS evaluation metrics.

Method	ClassFS ↓	ClusterFS ↓	MeanFS ↓
TSCV	0.645277	1.594577	1.119927
DSFG	0.750816	1.402442	1.076629
UFPT	0.362058	1.729917	1.045988
WVOP	1.219941	1.352989	1.286465
FTOP	1.141669	1.463282	1.302475
GLCM	1.028226	1.805302	1.416764
GLGK	1.234083	1.819762	1.526923
SGJP	0.388564	1.721218	1.054891
CATT	0.738440	1.530920	1.134680
FHKN	0.825380	1.520165	1.172773
DCSS (Proposed)	0.627423	1.454303	**1.040863**

VII. DISCUSSION

As mentioned above, this paper proposes a new model based on sentence similarity and deep clustering for topic detection which is based on "Azobel's cognitive theory of learning". The theory we are following in this article is that if the neural network first learns the concept of "sentence similarity", then the concept of "separating data for clustering", better results will be obtained. We named this theory as "Neural Network

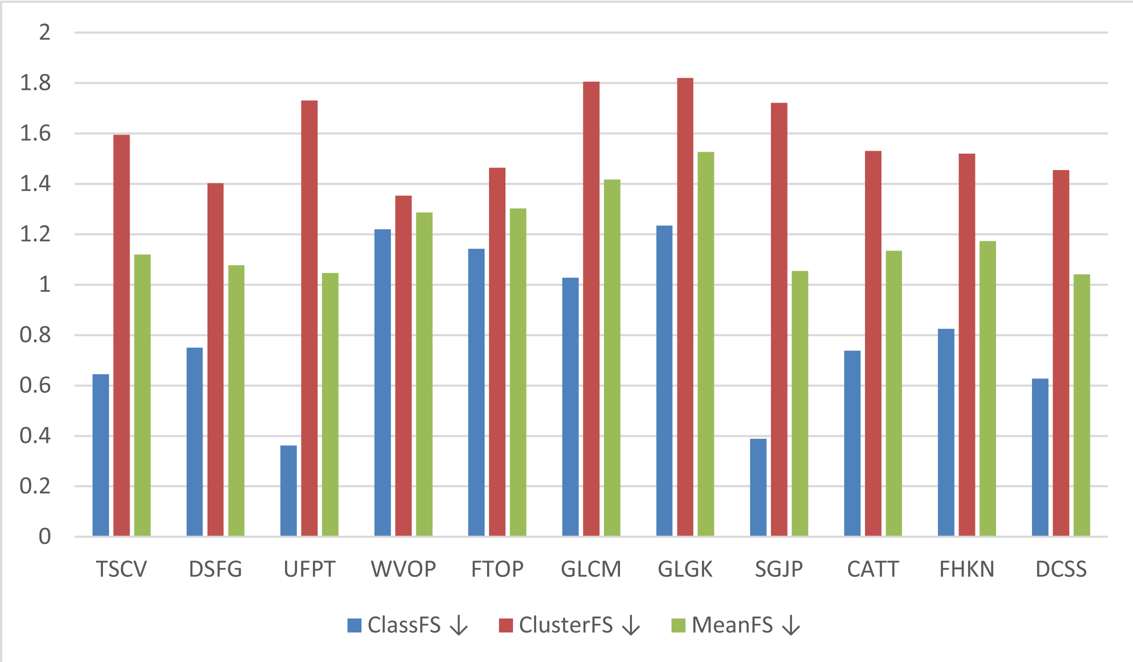

FIGURE 6. **Comparing the results of the proposed method (DCSS) with ten topic detection methods.**

Meaningful Learning'' (NNMeL) theory. This theory is similar to Azobel's theory about human. The first experiment attempts to test the correctness of this theory. The results of the first experiment show that the third strategy has achieved the best results in both tables. In the third strategy, the CSE and SSC components (which are involved in determining sentence similarity) are trained first. I.e. the network first learns the concept of ''sentence similarity''. Then, without changing the result of this training (i.e. without changing Contextualized Sentence Representation (CSR)) the concept of ''separating data for clustering'' is taught to the VSN part. Meanwhile, the generation of separable vectors is performed based on the CSR vectors i.e. the concept of ''separation'' is learned in connection with the concept of ''sentence similarity''. Therefore, the results of the first experiment confirm the concept of ''Neural Network Meaningful Learning''. This implies that meaningful learning in artificial neural network occurs when concepts are learned separately, but are related to each other.

As mentioned above, another novelty of this article is the generation of latent space vectors based on sentence similarity. The question is: how effective is this method? The second to fourth experiments aim to demonstrate the performance of ''Deep Clustering Based on Sentence Similarity'' (DCSS) is better than the existing methods. Since the basis of DCSS method is deep clustering, we first compared it with a state-of-the-art deep clustering method. We not only evaluated it using existing evaluation criteria (second test) but also assessed it manually (third test). The results from both experiments confirm that using DCSS and training based on sentence similarity is better than using autoencoder and training based on denoising. The fourth experiment, however, follows a different perspective. In this experiment, we evaluated the proposed model from the topic detection point of view instead of clustering point of view. The results of this experiment also confirm the performance of the proposed model.

VIII. CONCLUSION

In this paper, a new theory for neural network training named ''Neural Network Meaningful Learning Theory'' is presented. According to this theory, meaningful learning in an artificial neural network occurs when concepts are learned separately but are also related to each other. The application of this theory can be expressed as follows: if we want to train an ANN for an application, we should divide it into several concepts and assign a representation vector to each concept separately. Subsequently, each concept should be trained separately, one after the other, while making connections between them using their corresponding representation vectors. As a testbed for evaluating this theory, the DCSS model is proposed. The model is intended to learn two concepts: ''sentence similarity'' and ''data separation for clustering''. Therefore, in the proposed model, first the concept of ''sentence similarity'' is taught to the network, then the concept of ''data separation for clustering''. The connection between these two concepts is established through the CSR (contextualized sentence representation) vector. An experiment was also conducted to evaluate this theory. The results of the experiment confirmed that training according to this theory yields better results in compare with training outside of it. It also had another important conclusion for us: ''it is better to separate role of the vectors in the learning''. In the current researches same vector plays different roles for example presenting ''latent space'' and ''separable vectors of clustering''. The results of this study indicates that better results are achieved when different vectors are used for each role and trained through NNMeL theory.

In addition, a new model named ''Deep Clustering Based on Sentence Similarity'' (DCSS) is proposed for topic detection. The main idea of this model is ''to generate a representation of sentences based on their similarity'' as opposed to other methods of generating representations regardless of similarity. In addition, a framework is provided to make the model trainable. The conducted tests show the better performance of this model compared to the model that did not use sentence similarity. These tests were performed not only using NMI and Accuracy criteria, but also using human evaluation from clustering perspective. The results indicated a 6.6% improvement in accuracy. In addition, the proposed model was compared with ten other methods from topic detection point of view. The results of these tests also showed the superiority of the proposed model.

DCSS utilizes ''sentence similarity'' module instead of ''autoencoder'' module used in current methods. The autoencoder is simpler to train because it doesn't require any special dataset for training. A noise is simply added to the input vector and is used to train autoencoder. However, a similarity dataset is needed to train DCSS, which is a drawback of this model. In contrast, the results are superior and closely resemble human nature. Additionally, the sentence similarity module can be trained on a large dataset once and then used as a pretrained module for other applications.

IX. FUTURE WORKS

The findings and models presented in this paper open several avenues for future research and development. The most important one is to conduct empirical studies to validate the NNMeL theory by implementing it in various neural network architectures and comparing their performance with traditional training methods. This theory can undergo testing in various research domains and practical applications. Here, some examples of applications in which this theory can be applied are provided along with their base papers for those who want to work in this field: enhancing recommender systems leveraging the principles of [30] and [31], image processing [32], [33], question answering [34] and user reviews for products [35], [36].

REFERENCES

[1] Q. Xu, H. Gu, and S. Ji, ''Text clustering based on pre-trained models and autoencoders,'' *Frontiers Comput. Neurosci.*, vol. 17, pp. 1–13, Jan. 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fncom.2023.1334436/full, doi: 10.3389/fncom.2023.1334436.

[2] D. P. Ausubel, *The Psychology of Meaningful Verbal Learning*. New York, NY, USA: Grune & Stratton, 1963.

[3] D. P. Ausubel, *The Acquisition and Retention of Knowledge: A Cognitive View.* The Netherlands: Springer, 2000. [Online]. Available: https://www.google.com/books/edition/_/cwV_1uIpgVAC?hl=en&sa=X&ved=2ahUKEwjww-Lw-PWEAxXT0AIHHUbMCPEQ8fIDegQIFBAI

[4] T. G. K. Bryce and E. J. Blown, ''Ausubel's meaningful learning revisited,'' *Current Psychol.*, pp. 1–20, Apr. 2023. [Online]. Available: https://link.springer.com/article/10.1007/s12144-023-04440-4#citeas, doi: 10.1007/s12144-023-04440-4.

[5] V. C. Handa and V. M. Talisayon, ''Practical problem-solving tasks for meaningful learning and retention in college chemistry for pre-service teachers,'' *Eur. J. Sci. Math. Educ.*, vol. 11, no. 4, pp. 702–716, Oct. 2023.

[6] M. M. Abedin, ''The application of learning theories in language class: A scenario in the preschools in Dhaka city,'' *J. Bus., Arts Soc. Sci.*, vol. 1, no. 1, pp. 14–22, 2023.

[7] A. S. Magalhães and M. P. Nóbrega, *Statistical Literacy and Meaningful Learning: A Dialogical Counterpart*. Brazil, Switzerland: S. Editora, 2023, ch. 7. [Online]. Available: https://sevenpublicacoes.com.br/index.php/editora/article/view/2239

[8] C. Huang, Z. Han, M. Li, X. Wang, and W. Zhao, ''Sentiment evolution with interaction levels in blended learning environments: Using learning analytics and epistemic network analysis,'' *Australas. J. Educ. Technol.*, vol. 37, no. 2, pp. 81–95, May 2021, doi: 10.14742/ajet.6749.

[9] W. Zheng, G. Gong, J. Tian, S. Lu, R. Wang, Z. Yin, X. Li, and L. Yin, ''Design of a modified transformer architecture based on relative position coding,'' *Int. J. Comput. Intell. Syst.*, vol. 16, no. 1, pp. 1–17, Oct. 2023. [Online]. Available: https://link.springer.com/article/10.1007/s44196-023-00345-z, doi: 10.1007/s44196-023-00345-z.

[10] F. Atefeh and W. Khreich, ''A survey of techniques for event detection in Twitter,'' *Comput. Intell.*, vol. 31, no. 1, pp. 132–164, Feb. 2015, doi: 10.1111/coin.12017.

[11] R. Ibrahim, A. Elbagoury, M. S. Kamel, and F. Karray, ''Tools and approaches for topic detection from Twitter streams: Survey,'' *Knowl. Inf. Syst.*, vol. 54, no. 3, pp. 511–539, Mar. 2018, doi: 10.1007/s10115-017-1081-x.

[12] M. Hasan, M. A. Orgun, and R. Schwitter, ''A survey on real-time event detection from the Twitter data stream,'' *J. Inf. Sci.*, vol. 44, no. 4, pp. 443–463, Aug. 2018, doi: 10.1177/0165551517698564.

[13] Z. Mottaghinia, M.-R. Feizi-Derakhshi, L. Farzinvash, and P. Salehpour, ''A review of approaches for topic detection in Twitter,'' *J. Experim. Theor. Artif. Intell.*, vol. 33, no. 5, pp. 747–773, Jun. 2020, doi: 10.1080/0952813x.2020.1785019.

[14] M. Asgari-Chenaghlu, M.-R. Feizi-Derakhshi, L. Farzinvash, M.-A. Balafar, and C. Motamed, ''Topic detection and tracking techniques on Twitter: A systematic review,'' *Complexity*, vol. 2021, pp. 1–15, Jun. 2021, doi: 10.1155/2021/8833084.

[15] E. Zafarani-Moattar, M. R. Kangavari, and A. M. Rahmani, ''Topic detection on COVID-19 tweets: A comparative study on clustering and transfer learning models,'' *Tabriz J. Electr. Eng.*, vol. 52, no. 4, pp. 281–291, 2022, doi: 10.22034/tjee.2022.15744.

[16] L. Wang, C. Gao, J. Wei, W. Ma, R. Liu, and S. Vosoughi, ''An empirical survey of unsupervised text representation methods on Twitter data,'' in *Proc. 6th Workshop Noisy User-Generated Text (W-NUT)*, Dec. 2020, pp. 209–214, doi: 10.18653/v1/2020.wnut-1.27.

[17] J. Xu, B. Xu, P. Wang, S. Zheng, G. Tian, J. Zhao, and B. Xu, ''Self-taught convolutional neural networks for short text clustering,'' *Neural Netw.*, vol. 88, pp. 22–31, Apr. 2017, doi: 10.1016/j.neunet.2016.12.008.

[18] Y. Ren, J. Pu, Z. Yang, J. Xu, G. Li, X. Pu, P. S. Yu, and L. He, ''Deep clustering: A comprehensive survey,'' 2022, *arXiv:2210.04142*.

[19] Y. Bougteb, B. Ouhbi, B. Frikh, and E. M. Zemmouri, ''Deep learning based topics detection,'' in *Proc. 3rd Int. Conf. Intell. Comput. Data Sci. (ICDS)*, Oct. 2019, pp. 1–7, doi: 10.1109/ICDS47004.2019.8942245.

[20] H. Murfi, N. Rosaline, and N. Hariadi, ''Deep autoencoder-based fuzzy c-means for topic detection,'' *Array*, vol. 13, Mar. 2022, Art. no. 100124, doi: 10.1016/j.array.2021.100124.

[21] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long, ''A survey of clustering with deep learning: From the perspective of network architecture,'' *IEEE Access*, vol. 6, pp. 39501–39514, 2018, doi: 10.1109/ACCESS.2018.2855437.

[22] B. L. Ibtihel, H. Lobna, and B. R. Lotfi, ''A deep learning-based ranking approach for microblog retrieval,'' *Proc. Comput. Sci.*, vol. 159, pp. 352–362, Jan. 2019, doi: 10.1016/j.procs.2019.09.190.

[23] A. Hadifar, L. Sterckx, T. Demeester, and C. Develder, ''A self-training approach for short text clustering,'' in *Proc. 4th Workshop Represent. Learn.*, 2019, pp. 194–199, doi: 10.18653/v1/w19-4322.

[24] M. W. Akram, M. Salman, M. F. Bashir, S. M. S. Salman, T. R. Gadekallu, and A. R. Javed, "A novel deep auto-encoder based linguistics clustering model for social text," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, Apr. 2022, doi: 10.1145/3527838.

[25] P. Dahal, "Learning embedding space for clustering from deep representations," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 3747–3755, doi: 10.1109/BIGDATA.2018.8622629.

[26] O. Gencoglu, "Deep representation learning for clustering of health tweets," 2018, *arXiv:1901.00439*.

[27] S. Hosseini and Z. A. Varzaneh, "Deep text clustering using stacked autoencoder," *Multimedia Tools Appl.*, vol. 81, no. 8, pp. 10861–10881, Mar. 2022, doi: 10.1007/s11042-022-12155-0.

[28] M. Ranjbar-Khadivi, M. R. Feizi-Derakhshi, A. Forouzandeh, P. Gholami, A. R. Feizi-Derakhshi, and E. Zafarani-Moattar, (2022), "Sep_TD_Tel01," *Mendeley Data*, doi: 10.17632/372rnwf9pc.

[29] E. Zafarani-Moattar, M. R. Kangavari, and A. M. Rahmani, "A comprehensive study on frequent pattern mining and clustering categories for topic detection in Persian text stream," 2024, *arXiv:2403.10237*.

[30] Y. Xu, E. Wang, Y. Yang, and Y. Chang, "A unified collaborative representation learning for neural-network based recommender systems," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 11, pp. 5126–5139, Nov. 2022, doi: 10.1109/TKDE.2021.3054782.

[31] F. Huang, Z. Wang, X. Huang, Y. Qian, Z. Li, and H. Chen, "Aligning distillation for cold-start item recommendation," in *Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2023, pp. 1147–1157, doi: 10.1145/3539618.3591732.

[32] Y. Wang, N. Xu, A.-A. Liu, W. Li, and Y. Zhang, "High-order interaction learning for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4417–4430, Jul. 2022, doi: 10.1109/TCSVT.2021.3121062.

[33] C. Fu, H. Yuan, H. Xu, H. Zhang, and L. Shen, "TMSO-Net: Texture adaptive multi-scale observation for light field image depth estimation," *J. Vis. Commun. Image Represent.*, vol. 90, Feb. 2023, Art. no. 103731, doi: 10.1016/j.jvcir.2022.103731.

[34] Y. Liu, G. Li, and L. Lin, "Cross-modal causal relational reasoning for event-level visual question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 11624–11641, Oct. 2023, doi: 10.1109/TPAMI.2023.3284038.

[35] Z. Liu et al., "Emotion-semantic-aware dual contrastive learning for epistemic emotion identification of learner-generated reviews in MOOCs," *IEEE Trans. Neural Netw. Learn. Syst.*, doi: 10.1109/TNNLS.2023.3294636.

[36] Z. Wu, J. Cao, Y. Wang, Y. Wang, L. Zhang, and J. Wu, "HPSD: A hybrid PU-learning-based spammer detection model for product reviews," *IEEE Trans. Cybern.*, vol. 50, no. 4, pp. 1595–1606, Apr. 2020.

ELNAZ ZAFARANI-MOATTAR received the B.S. degree in software engineering from Islamic Azad University, Tabriz Branch, and the M.S. degree in artificial intelligence from Islamic Azad University, Science and Research Branch, where she is currently pursuing the Ph.D. degree in artificial intelligence. She is also a Faculty Member of Islamic Azad University, Tabriz Branch. Her research interests include natural language processing, social network analysis, and deep learning.

MOHAMMAD REZA KANGAVARI received the B.Sc. degree in computer science from the Sharif University of Technology, the M.Sc. degree from Salford University, and the Ph.D. degree from The University of Manchester. He is currently an Associate Professor with Iran University of Science and Technology. His research interests include intelligent systems, human–computer-interaction, cognitive computing, machine learning, and sensor networks.

AMIR MASOUD RAHMANI received the B.S. degree in computer engineering from Amir Kabir University, Tehran, in 1996, the M.S. degree in computer engineering from the Sharif University of Technology, Tehran, in 1998, and the Ph.D. degree in computer engineering from IAU University, Tehran, in 2005. He is currently a Professor in computer engineering. His research interests include distributed systems, the Internet of Things, and evolutionary computing.