IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 August 2023, accepted 4 September 2023, date of publication 6 September 2023, date of current version 12 September 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3312622*

RESEARCH ARTICLE

# A Novel Two-Fold Loss Function for Data Clustering and Reconstruction: Application to Document Analysis

**MEBARKA ALLAOUI[1,2], (Member, IEEE), MOHAMMED LAMINE KHERFI[3,4], OUSSAMA AIADI[1,2], AND SAMIR BRAHIM BELHAOUARI[5], (Senior Member, IEEE)**

[1]Department of Computer Science and Information Technologies, University Kasdi Merbah Ouargla (UKMO), Ouargla 30000, Algeria
[2]Artificial Intelligence and Information Technology Laboratory (LINATI), University Kasdi Merbah (UKMO), Ouargla 30000, Algeria
[3]National Higher School of Artificial Intelligence, Algiers, Algeria
[4]LAMIA Laboratory, Université du Québec à Trois-Rivières (UQTR), Trois-Rivières, QC G9A 5H7, Canada
[5]Division of Information and Computing Technology, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

Corresponding authors: Mebarka Allaoui (allaoui.mebarka@univ-ouargla.dz) and Samir Brahim Belhaouari (sbelhaouari@hbku.edu.qa)

**ABSTRACT** In the midst of the ongoing COVID-19 pandemic, there has been a surge in scientific literature aimed at understanding the virus and its impact. However, it has become challenging for a researcher to deal with thousands of articles published daily. This paper proposes a novel deep-learning architecture to organize a large dataset of COVID-19-related scientific literature and provides a clear overview of the current state of knowledge. The proposed model is developed based on two main bases to ensure robustness and efficiency. In particular, we trained a denoising autoencoder with clean and noisy data to make the model can balance, preserving the underline structure and generalizing the new unseen data. Furthermore, the cornerstone of the proposed architecture lies in training the autoencoder using a two-fold objective function that jointly incorporates the data's reconstruction and clustering. The advantage behind this combination is to avoid the distortion of the latent space and to improve the model efficiency. Afterward, we use the Latent Dirichlet Allocation (LDA) to analyze the document's topics. For the sake of computational efficiency, instead of feeding the LDA with the whole dataset of documents, we fed it with the clusters produced in the phase of dimensionality reduction and clustering to count the frequency of topics in each cluster. The model was trained on a large public corpus of COVID-19-related articles and evaluated using a set of evaluation metrics. Experimental results indicate the superiority of our proposed model compared to several recent studies.

**INDEX TERMS** Clustering, COVID-19, deep learning, dimensionality reduction, document organization, topic modeling.

## I. INTRODUCTION

The COVID-19 pandemic [1], which broke out in 2019 and continues to do so, poses a significant threat to humanity, resulting in hundreds of millions of cases and a few million deaths [2], [3]. COVID-19 has swept through the globe and affected the health system, causing unparalleled economic downturn, the total quarantine of citizens, the closing of international borders, and the unavailability of many services [4], [5].

All governments have made significant efforts to stop the epidemic from spreading and limit the damage it causes. In addition, thousands of researchers worldwide were and still are attempting, in their respective disciplines, to produce new vaccines, re-purpose or propose new medications, develop means for tracking population contamination, and research the impacts of the lockdown on countries' economies and individual psyches, and so on [6], [7], [8], [9]. This has resulted in a rapid increase in the number of scientific articles covering COVID-19, SARS-CoV-2, and related coronaviruses [10]. The number of published papers is over 1,000,000 scholarly articles, including over 400,000 with full

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

text, which continues to increase exponentially. COVID-19-related documents cover a wide range of disciplines and have specific requirements with specific characteristics, such as the type of data (e.g., clinical data, epidemiological data) that require particular processing and analysis.

Indeed, staying up to date with the latest scientific advances in this field, navigating through the hundreds of thousands of related articles that were published recently, or finding articles that answer a specific question or deal with a particular aspect of COVID-19 has become extremely difficult, for a researcher or decision-maker. Locating the desired COVID-19 documents within this huge corpus of documents is important for several reasons. First, each document handles the COVID-19 issue from a different point of view depending on the field in which the authors work. For instance, economists try to estimate the economic losses caused by the pandemic and to which extent the economy can resist this issue. In addition, data scientists are interested in incorporating AI techniques to facilitate the detection of this disease. This richness and interdisciplinary motivates the development of efficient tools to organize the existing documents and to reach the desirable documents effectively. Second, if one is interested in AI-based solutions for COVID-19, picking up the appropriate AI technique to deploy depends on several factors. For instance, the performance of AI solutions can be heavily influenced by the imaging modality to consider (e.g., Computed Tomography scan or Chest X-Ray). Another essential question to answer before adopting a specific technique is to check whether the technical requirements are satisfied. For instance, compared to the conventional approach, deep-learning-based solutions require a large amount of data to train and hardware with high specifications.

Recently, very few studies have discussed this topic. From one point of view, the existing studies can be categorized into conventional and deep-based approaches. The first category includes the works [11], [12], [13] that use the conventional machine learning approaches involving Principal Component Analysis, Expectation Maximization (EM), t-distributed student neighbor embedding (t-SNE), etc. The second approach involves developing deep-based methods to organize COVID-19-related literature [14]. Although these studies have dealt with the problem at hand, much effort is still needed to achieve the desired goal. In particular, the conventional machine learning techniques [12] lack the generalization power, especially with the constantly increasing number of documents. For instance, considering the k-means algorithm to perform document clustering can yield inconsistent solutions because of the high dimension of data. In addition, for the deep architectures based on auto-encoder (AE) [15], the AE is designed to perform both document clustering and input reconstruction. Nevertheless, performing such joint processes separately could negatively affect the learned latent space.

This paper introduces a novel deep learning-based framework called Deep Joint Reconstruction and clustering (DJRC) for exploring and organizing a large-scale dataset of COVID-19-related documents. The proposed method jointly performs dimensionality reduction and clustering using a two-fold objective function. More precisely, we propose an architecture of denoising autoencoder with three components: two different encoders and one decoder. The two encoders are called clean and noisy encoders, depending on the data type they are trained on (clean and noisy data, respectively). Noisy data are documents that are not pre-processed, while clean documents are pre-processed prior to feeding them to the auto-encoder. The three components are trained using the two-fold objective function, where the first term in this function represents the reconstruction loss, which aims to reduce the input features' dimensionality. The second term seeks to predict cluster assignments. Two distinct matrices associated with clean and noisy encoders during the training phase are updated iteratively to accomplish this. Afterward, the Latent Dirichlet Allocation (LDA) is employed for topic modeling. LDA identifies the main topics covered in the documents to understand the content better and analyze the relationships in the data.

To sum up, the contributions of the current study can be summarized as follows:

- We introduce a novel robust deep architecture for COVID-19-related document analysis.
- To ensure faithful learning, we propose a novel two-fold objective function to train the proposed architecture, which simultaneously performs dimensionality reduction and clustering assignments.
- The proposed method can achieve high robustness due to considering both clean and noisy encoders. The noisy encoder calculates the reconstruction loss, and clean and noisy latent spaces are used for clustering assignments.
- For the sake of computational efficiency, instead of feeding the LDA by the whole dataset of documents that cover different fields, e.g., economy, AI, biology,... etc., we fed it by the clusters (which group documents of the same field) predicted by the model.
- We conduct thorough experiments to evaluate the performance of the proposed approach. The experimental results demonstrate the effectiveness of our approach, which significantly outperformed several recent studies.

The rest of this paper is organized as follows: Section II reviews related work. More details about our method are in Sec. III. Section IV reports some experimental results. Then, we conclude the paper and give some of our perspectives for future research in Sec. V.

## II. RELATED WORK

Machine Learning (ML) has been successfully applied to solve many problems in different fields [16], [17], [18]. Natural language processing (NLP) is one of the ML fields which involves the engineering of computational models and processes to solve practical problems in understanding human languages. Documents organization and clustering, which is the main target of this study, is a typical task of NLP. In the past decades, machine learning approaches

such as naïve Bayes [19], k-nearest neighbors [20], hidden Markov models [21], support vector machines [22] were widely used for documents organization. However, with the impressive performance of deep learning on different applications, deep models involving convolutional neural networks (CNNs) [23], [24], Recursive Neural Networks [25], have widely been considered to handle different NLP tasks. This study is particularly interested in the automatic organization of the COVID-19 literature documents. It is worth mentioning that a few literature studies are concerned with such an interesting topic. Existing studies can be classified into two categories, depending on the techniques used: Conventional and Deep-based approaches.

The first category involves methods that use conventional machine learning techniques such as expectation maximization (EM) algorithm and principal component analysis (PCA). For instance, in [11], the different foci of many COVID-19-related abstracts were analyzed using a clustering approach, where Singular Value Decomposition (SVD) was used for dimensionality reduction, and the EM algorithm was employed to perform clustering. Nevertheless, this study was concerned with examining only the abstracts and not considering the entire document. This could negatively affect the determination of topics discussed in these documents. Authors in [12] have proposed COVID-19 LC, an organization and visualization tool for COVID-19 documents. At first, documents are pre-processed and transformed into vectors using the TF-IDF algorithm. Then, Principal Component Analysis (PCA) is used for dimensionality reduction. This last step is a preliminary step for the next step in which t-Distributed Stochastic Neighbor Embedding (t-SNE) is used to visualize documents. Next, the unsupervised K-means cluster the documents by grouping similar data instances in the same cluster and entirely dissimilar from those in other clusters. Finally, Latent Dirichlet Allocation (LDA) is implemented to label the clusters. However, relying on K-means for clustering may arise several issues. For instance, it is well-known that K-means are unsuitable for databases with unbalanced clusters, such as the CORD-19 dataset. In addition, k-means is sensitive to the initial clustering solution and cannot appropriately handle outliers and noisy data. In [13], authors built a platform to extract information on COVID-19 clinical risk variables and present the results clustered to aid knowledge discovery. The authors conducted a comparative study between two clustering algorithms which are spectral and Agglomerative clustering. There is still room for improvement, especially in reducing the dimensionality of vectors extracted from documents. Additionally, improving the clustering component could significantly improve the outcomes of the proposed model [26], [27].

As for the second category (i.e., deep learning-based methods), studies using deep learning for COVID-19-related document analysis are very scarce. In [14], the authors proposed a method for organizing and visualizing COVID-19-related documents from the CORD-19 dataset. After the pre-processing step, the dimensions of the dataset were reduced using a deep-stacked autoencoder. Later, the reduced dataset was projected into a 2D space using Uniform Manifold Approximation and Projection (UMAP) [28], and the agglomerative clustering algorithm was used to cluster the data. Finally, the topic modeling step was performed using the LDA.

As for the other deep learning methods [15] that are dedicated to clustering, they suffer from several limitations. First, in [15], clustering and latent space learning (i.e., dimensionality reduction) are done separately, i.e., reconstruction loss is optimized. The decoder is removed to train the model to perform the clustering assignments. However, in such a manner, the learned latent space could be deformed [29], significantly affecting the clustering process and increasing the time required to do this task. Second, considering either clean or noisy encoders can reduce the generalization power of the model. Moreover, [15] is based on computing the similarity between clustering assignment matrices using the KL-divergence. In [28], it has been shown that KL-D considers only one direction (i.e., from the first matrix to the second one) to generate the similarity. This can cause loss of significant information compared to the other metrics (e.g., cross-entropy), which considers the two directions in computing the similarity between the matrices.

## III. OUR METHOD

This section is devoted to presenting the proposed method in detail. Indeed, the main target of this study is to develop a clustering method to organize the scientific documents related to COVID-19 based on deep learning. To do so, we propose a novel deep architecture that simultaneously carries out dimensionality reduction and data clustering (DJRC). Figure 1 presents the general flowchart of the proposed approach. From Figure 1, we can notice that the first step of our method is to pre-process the input documents by eliminating punctuation, stop words, orthographic and spelling errors, symbols, etc. We consider an autoencoder-based architecture in which three components, two different encoders and one decoder, are jointly trained. We refer to the two encoders as clean and noisy because the former is fed with pre-processed documents, whereas the last one is fed with raw documents. To improve the model robustness, we consider training the noisy encoder with the weights of the clean encoder. To further improve the model robustness, inspired by the denoising auto-encoder (DAE) principle, we consider a two-fold loss (i.e., objective function), where the first fold of the loss is generated by considering the latent space of the noisy encoder and the output of the decoder. The second fold of the loss is designed for predicting the cluster assignments. To do so, we associate each clean and noisy encoder with two matrices, aiming to converge the matrices to each other by iteratively updating them during training. This double-side convergence can be achieved by considering the cross-entropy function instead of the conventional KL divergence. Indeed, this two-fold loss (i.e., objective function) allows us to reduce the dimensionality and cluster the

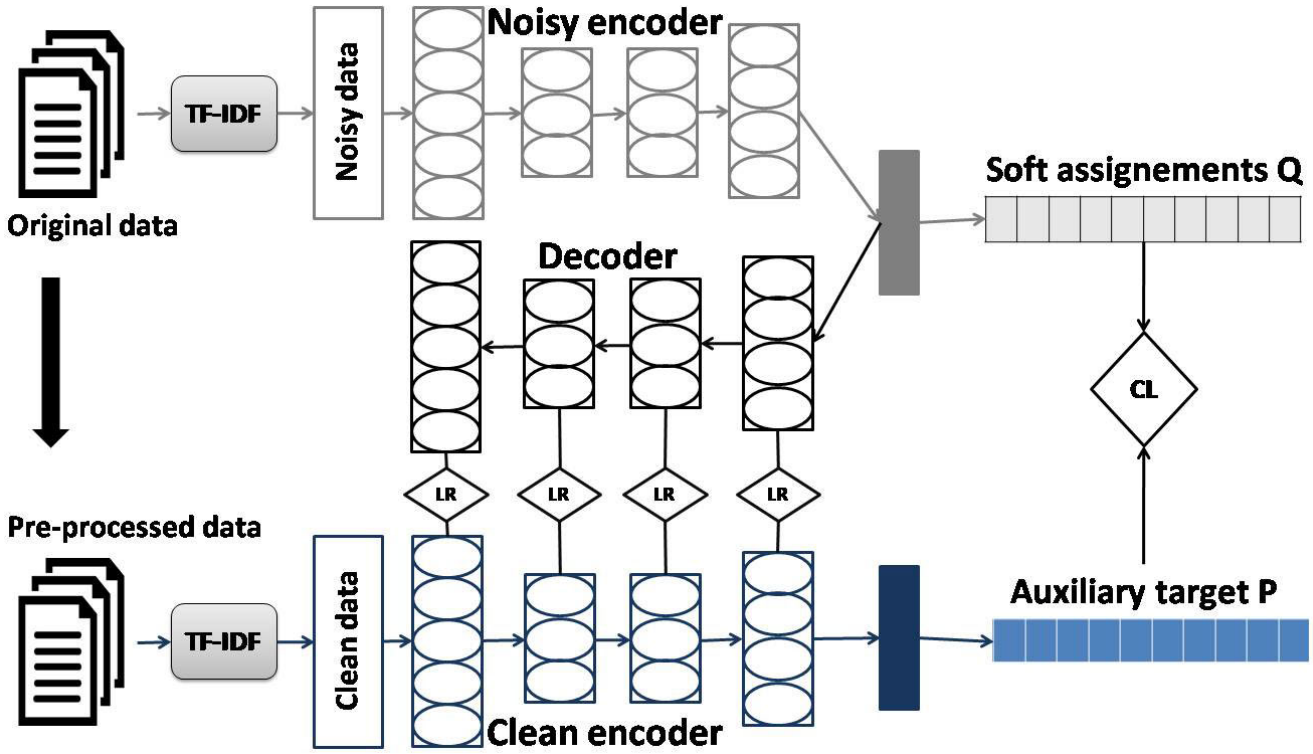



**FIGURE 1.** DJRC's architecture. The clean and noisy data are passed to dimensionality reduction and clustering algorithms to embed and cluster the dataset.

data simultaneously. The cornerstone of the proposed method lies in this simultaneity, which allows our model to train efficiently and overcome the issue of latent space distortion.

After the training process had finished, and unlike the existing works, which fed the entire dataset documents to the topic modeling model, we fed the LDA with the data clusters, which are supposed to be consistent and group the documents covering the same topic, as shown in Figure 2. Doing so can significantly improve the model efficiency and reduce the response time, which is crucial for such models and for fighting COVID-19 in general.

### A. DATASET PRE-PROCESSING

In our work, we use the COVID-19 Open Research Dataset (CORD-19) [30], which is a freely and publicly available dataset on the Kaggle website. It contains over 1,000,000 scholarly articles, including over 400,000 full-text scientific papers concerning COVID-19, SARS-CoV2, and other related coronaviruses. In the natural language processing field, the pre-processing step (such as removing the punctuation, stop-word, etc.) is essential to filter and clean the data from inaccuracies, errors, or conflicting information to improve the performance of the proposed model [31]. Furthermore, we used the well-known Term Frequency-inverse document frequency (Tf-IDF) algorithm [32] to convert the documents from the text format to the feature vectors that can be processed by learning models later on.

TF-IDF vectorization involves calculating the TF-IDF score for every word in the corpus relative to that document and then putting that information into a vector. Thus each document in the corpus would have its vector, and the vector would have a TF-IDF score for every single word in the entire collection of documents. Then the similarity of the documents can be computed using cosine similarity between the vectors of those documents.

### B. DIMENSIONALITY REDUCTION AND CLUSTER ASSGINMENTS

#### 1) THE PROPOSED ARCHITECTURE

In our model, Deep Joint dimensionality reduction and clustering (DJRC) is based on a Denoising Autoencoder (DAE) with three components. DAE is a good dimensionality reduction technique that can extract the dataset's intrinsic structure [33]. Unlike the standard learning approach for denoising autoencoders, we built an autoencoder similar to the one reported in [34]. Figure 1 (a) represents the structure of the DAE, which contains three parts: a clean encoder, a noisy encoder, and a decoder part. The clean encoder is used to compute the more accurate target variables, while the noisy encoder is trained to achieve noise-invariant predictions. The clean and noisy encoders are trained together with the decoder, where the clean encoder shares its weights with the noisy encoder. In the following, we give more details on these deep architectures.

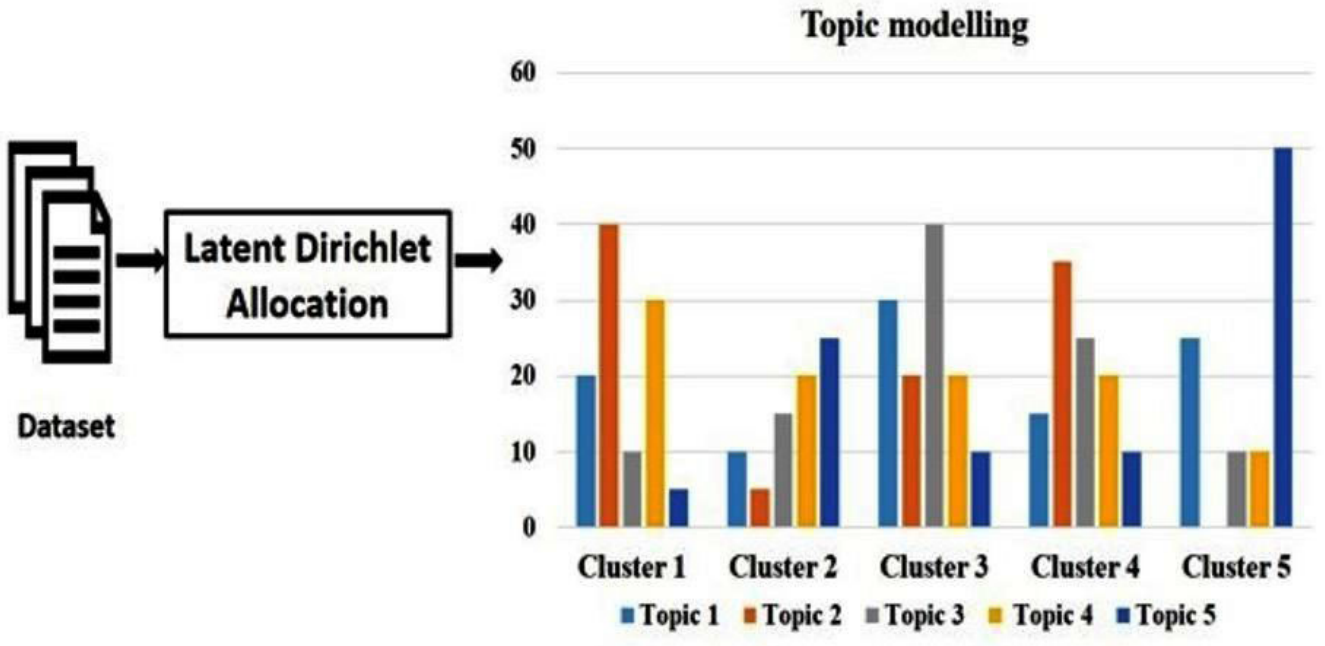



**FIGURE 2.** Topic Modeling step. We use LDA to model the topic and extract the keywords that best represent each cluster topic.

### a: NOISY ENCODER

This component is trained using noisy input data. Noisy data refers to text containing errors, inconsistencies, irrelevant information, punctuation, stop-word, numbers, special characters, capitalization, etc. The aim of training an autoencoder using noisy data is to make it more robust to variations in the input data and to increase its ability to generalize to new, unseen data. When the autoencoder is trained on noisy data, it must learn to reconstruct the original, clean data from a corrupted version. This helps the model to become more robust to noise and to ignore irrelevant details in the input data. The following equation indicates the output of each layer in the noisy encoder:

$$\tilde{z}^l = f_e^l(W_e^l \tilde{z}^{(l-1)} + b_e^l) \tag{1}$$

$$\tilde{z}_l \sim Dropout(\tilde{z}_l) \tag{2}$$

where $\tilde{z}^l$ is the noisy input of the l-th encoding layer. Dropout(.) is a regularization technique that we used to reduce the over-fitting in our autoencoder. $f_e^l$ is the activation function (Rectified Linear Unit (RELU) in our case) of the current encoding layer, and $\theta_e^l = \{W_e^l, b_e^l\}$ are the parameters of the l-th encoding layer, where $W$ stands for the weights, and $b$ represents the bias. The structure of the noisy encoder is $[D - 500 - 500 - 2000 - d]$, where $D$ is the dimension of the input data, and $d$ is the dimension of the latent space.

The noisy encoder is associated with a Soft assignment matrix denoted by Q. The Soft assignment matrix was computed using an equation similar to the one used in [28]. This equation measures the similarity between embedded point $\tilde{z}$ and centroid $\mu_k$:

$$q_{ik} = (1 + d_{ik}^2)^{-1} \tag{3}$$

where $d_{ik}^2$ is the squared distance between the data points $\tilde{z}$ and the centroids $\mu_k$, $d_{ik} = (\tilde{z} - \mu_k)$. $\tilde{z} = f(x_i) \in Z$ corresponds to the input data $x_i \in X$ after embedding.

The complexity of the noisy encoder is $O(N * W * e + Q)$, where $N$ is the number of samples, $W$ is the weights of the noisy encoder, and $e$ is the number of epochs.

### b: CLEAN ENCODER

The clean encoder is trained using data cleaned from noise or corruption that misleads the learning process. The aim of training an autoencoder using clean data is to learn a representation of the input data that is as accurate as possible. When the autoencoder is trained on clean data, it can learn the underlying structure of the data, including its important features and patterns. This allows the model to capture the input data's essence and accurately reconstruct the original data from its internal representation. The features of the clean encoder are used in the reconstruction loss function, which is inferred using the following equation.

$$z^l = f_e^l(W_e^l z^{(l-1)} + b_e^l) \tag{4}$$

where $z^l$ is the input of the l-th encoding layer. $f_e^l$ is the activation function of the current encoding layer, and $\theta_e^l = \{W_e^l, b_e^l\}$ are the parameters of the l-th encoding layer. The structure of the clean encoder is similar to the noisy encoder $[D - 500 - 500 - 2000 - d]$.

This component (i.e., the clean encoder) is associated with a matrix (denoted by P) which is iteratively updated during

training epochs to achieve the cluster assignments. $P$ represents the probabilistic point-to-cluster assignments using the obtained Soft assignment $Q$. Therefore, $P$ is the auxiliary target distribution proposed to improve feature representation and clustering assignment.

$$p_{ik} = \frac{q_{ik}/\sum_l q_{lk}}{\sum_m (q_{lm}/\sum_l q_{lm})} \tag{5}$$

$m$ is the number of clusters, and $l$ is the number of data points in the corresponding cluster.

The complexity of the clean encoder is $O(N * W * e + P)$, where $N$ is the number of samples, $W$ is the weights of the clean encoder, and $e$ is the number of epochs.

#### c: DECODER

The decoder is a commune part between the noisy and the clean encoders. By training on clean and noisy data, the decoder can learn a more robust representation of the data and generalize better to new, unseen data. The following equation represents the layers of the decoder part:

$$\hat{z}^l = f_d^l(W_d^l \hat{z} + b_d^l) \tag{6}$$

where $z_l$ is the input of the l-th decoding layer, $f_{d_l}$ is the activation function of the current decoding layer, and $\theta_d^l = \{W_d^l, b_d^l\}$ are the parameters of the i-th decoding layer. The structure of the decoder is $[d - 2000 - 500 - 500 - D]$, where $D$ is the dimension of the input data, and $d$ is the dimension of the latent space. The complexity of the decoder is $O(N * W * e)$, where $N$ is the number of samples, $W$ is the weights of the decoder, and $e$ is the number of epochs.

### 2) JOINT OPTIMIZATION USING THE TWO-FOLD OBJECTIVE FUNCTION

The bottleneck of the proposed algorithm is to provide a joint learning framework that optimizes the binary cross-entropy and autoencoder parameters. This is achieved by using a two-fold function, which is given by

$$\begin{aligned} min \sum_i \sum_l \|z_i^l - \hat{z}_i^l\|_2^2 \\ + \sum_i \sum_k p_{ik} \log \frac{p_{ik}}{q_{ik}} - (1 - p_{ik}) \log \frac{1 - p_{ik}}{1 - q_{ik}} \end{aligned} \tag{7}$$

Training the autoencoder using reconstruction and clustering losses simultaneously ensures the achievement of both objectives appropriately. The reconstruction loss helps to ensure that the autoencoder can effectively capture the underlying structure of the data and reduce it to a lower-dimensional representation. The clustering loss helps to force the encoder to produce a compact and well-separated representation of the data in the latent space, making it easier to perform the clustering task. The first term in Eq. (7) is the Reconstruction Loss (RL):

$$RL = \sum_i \sum_l \|z_i^l - \hat{z}_i^l\|_2^2 \tag{8}$$

It is worth mentioning that the reconstruction loss is obtained by considering the weights from the clean encoder and the features (i.e., of the latent space) from the noisy encoder. Combining the noisy encoder features and the clean encoder weights optimizes the reconstruction loss in a way that ensures the balance of the trade-off between better handling of the input data variations and the accuracy of the model.

The second term is the Clustering Loss (CL), computed between the Soft assignment matrix $Q$ and the auxiliary target matrix $P$. The CL is optimized until the Soft assignment matrix approximates the auxiliary target matrix. The CL term is defined as

$$CL(P||Q) = \sum_i \sum_k p_{ik} \log \frac{p_{ik}}{q_{ik}} - (1 - p_{ik}) \log \frac{1 - p_{ik}}{1 - q_{ik}} \tag{9}$$

We use the binary cross-entropy function as a clustering loss to perform double-side convergence, i.e., P converges to Q, and Q converges to P, which is inspired by UMAP. The Clustering Loss ($CL$) computes the total entropy between the quantity matrices $Q$ and $P$.

The aim is to find the derivative of $CL$ function w.r.t $z_i$ and $\mu_k$. First, $CL$ was noted as a function having $z_1, z_2, \ldots, z_n, \mu_1, \mu_2, \ldots, \mu_C$ as vector variables. Then, the data point coordinates $z_i$ are updated using the following iterative relations:

$$\begin{aligned} z_i^{(t+1)} = z_i^{(t)} - \eta \Delta_{z_i} CL(z_1^{(t)}, z_2^{(t)}, \ldots, z_i^{(t)}, \ldots, z_N^{(t)}, \\ \mu_1^{(t)}, \mu_2^{(t)}, \ldots, \mu_k^{(t)}, \ldots, \mu_C^{(t)}), \end{aligned} \tag{10}$$

for $i = 1, \ldots, N$ and $t$ is the time step. Where $\eta$ is the learning rate, $\Delta_{z_i} CL$ is the gradient of $CL$ loss function with respect to the components of the vector $z_i$ i.e., if $z_i = (z_{i1}, z_{i2}, \ldots, z_{id})^T$ then $\Delta_{z_i} CL = (\frac{\partial CL}{\partial z_{i1}}, \frac{\partial CL}{\partial z_{i2}}, \ldots, \frac{\partial CL}{\partial z_{id}})^T$ where $d = 1, \ldots, d$ is the dimension of data space and the subscript $T$ designate the transpose of the vector. The partial derivative of the CL loss function w.r.t each component d of the vector $z_i$ is:

$$\frac{\delta CL}{\delta z_i} = \sum_k \left[ \frac{2p_{ik}}{1 + d_{ik}^2} - \frac{2(1 - p_{ik})}{d_{ik}^2 (1 + d_{ik}^2)} \right] (z_i - \mu_k) \tag{11}$$

Similarly, the coordinates of the centers of the clusters $\mu_k$ are updated using the following iterative relations:

$$\begin{aligned} \mu_k^{(t+1)} = \mu_k^{(t)} - \eta \Delta_{\mu_k} CL(z_1^{(t)}, z_2^{(t)}, \ldots, z_i^{(t)}, \ldots, z_N^{(t)}, \\ \mu_1^{(t)}, \mu_2^{(t)}, \ldots, \mu_k^{(t)}, \ldots, \mu_C^{(t)}), \end{aligned} \tag{12}$$

for $k = 1, \ldots, C$ and $t$ is the time step. Where $\eta$ is the learning rate, $\Delta_{\mu_k} CL$ is the gradient of $CL$ loss function with respect to the components of the vector $\mu_k$ i.e., if $\mu_k = (\mu_{k1}, \mu_{k2}, \ldots, \mu_{kd})^T$ then $\Delta_{\mu_k} CL = (\frac{\partial CL}{\partial \mu_{k1}}, \frac{\partial CL}{\partial \mu_{k2}}, \ldots, \frac{\partial CL}{\partial \mu_{kd}})^T$. The partial derivative of the CL loss function w.r.t each component d of the vector $\mu_k$ is:

$$\frac{\delta CL}{\delta \mu_k} = \sum_k \left[ \frac{-2p_{ik}}{1 + d_{ik}^2} + \frac{2(1 - p_{ik})}{d_{ik}^2 (1 + d_{ik}^2)} \right] (z_i - \mu_k) \tag{13}$$

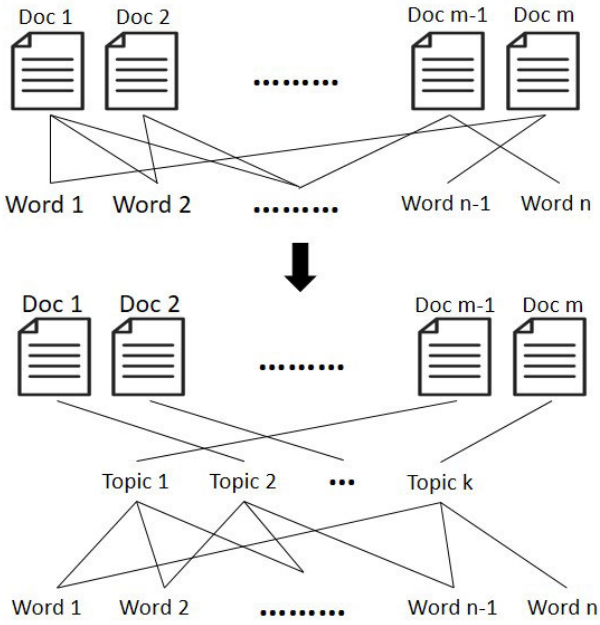



**FIGURE 3.** Latent Dirichlet allocation description.

Stochastic Gradient Descent (SGD) was used to optimize the cluster centers $\mu_k$ and the parameters of the DAE jointly. For each iteration, the gradients $\frac{\delta CL}{\delta z_i}$ are passed down to the deep autoencoder and update mapping parameters using back-propagation for computing $z_i$.

### C. TOPIC MODELING

In order to find the topics contained in the extracted clusters through our model to interpret and understand the data and gain insights into the reasons behind the model's decisions, we resorted to the topic modelling step. We used Latent Dirichlet Allocation (LDA) as a topic modelling technique. Unlike the previous studies, and to improve the computational efficiency of the proposed approach, we opt for using the clusters produced by the previous step (i.e., dimensionality reduction and clustering assignment). These clusters are fed to the LDA instead of the whole dataset. LDA [28] is a widely used method in various applications. In LDA, each document is described by a distribution of topics, and the distribution of words can describe each topic, as shown by Figure 3.

Many reasons make LDA a good choice as a topic modelling technique. LDA is flexible, effective, and scalable to handle large datasets such as CORD-19. LDA is a generative model that explicitly models the generation of documents based on a set of latent topics. This makes it easy to interpret the model's results and understand the relationships between topics and documents. LDA is an unsupervised learning technique that does not require labelled data, which makes it useful in our case.

## IV. EXPERIMENTAL RESULTS

In this section, we report our experimental findings. Specifically, we carry out several experiments to measure the performance of our proposed architecture. We devid these experiments into two studies: comparative study and case study as follows:

- In the Comparative study, we compare the performance our model against several baseline models on benchmark datasets.
- In the case study, we study the performance of our model on important topic, which is the documents related to COVID-19.

Let us first represent the evaluation metrics, the dataset we considered in these experiments, and the implementation details.

### A. EXPERIMENTAL SETUP

#### 1) DATASET

We conduct experiments on four benchmark datasets given as follows:

1) **USPS**: consists of 9298 images belonging to 10 different classes. Each image is a 16 × 16 grey-scale image [35].
2) **MNIST-FULL**: consists of 10 handwritten digits with 70,000 images. Each image is a 28 × 28 grey-scale image [36].
3) **MNIST-Test**: consists of 10 handwritten digits with 10,000 images. Each image is a 28 × 28 grey-scale image.
4) **The COVID-19 Open Research Dataset Challenge (CORD-19)**: is a freely available dataset of more than 570,000 scientific papers about COVID-19, the virus SARS-CoV2 that caused it, and other related coronaviruses. Among those papers, over 150,000 are brought with their full text [30].

#### 2) EVALUATION METRICS

We adopt two external metrics and two internal metrics to evaluate the clustering task and one metric for topic modeling.

The two external clustering metrics are Accuracy (ACC) and Normalized Mutual Information (NMI):

$$ACC = max_m \frac{\sum_{i=1}^{n} 1\{y_i = m(c_i)\}}{n} \tag{14}$$

where $y_i$ and $c_i$ are the ground truth and predicted label of sample $i$ respectively. All conceivable one-to-one mappings between clusters and labels are covered by the range $m$.

$$NMI = \frac{2I(y, c)}{[H(y) + H(c)]} \tag{15}$$

where $y$ is the ground truth label, $c$ is the predicted label, H(.) is the Entropy and $I(y, c)$ is the Mutual Information between $y$ and $c$.

The two internal clustering metrics are:

- **Davies-Bouldin (DB)** [37]: which takes into account both intra-dispersion of clusters and their inter-structure. Therefore, the performance of the clustering algorithm is better when the DB index is near zero.

- **Silhouette Coefficient (SC)** [38]: SC measures the goodness of clustering. The performance of the clustering algorithm is better when the SC index is near 1, and it will be worse when the SC index is near −1.

To evaluate the performance of LDA, we use the following metric:

- **Coherence Score** $C_v$: It calculates how often two words, $w_i$ and $w_j$ appear together in the corpus, and it's defined as:

$$C_{UMass}(w_i, w_j) = \log \frac{D(w_i, w_j) + 1}{D(w_i)},$$

where $D(w_i, w_j)$ indicates how many times words $w_i$ and $w_j$ appear together in documents, and $D(w_i)$ is how many time word $w_i$ appeared alone. The greater the number, the better is coherence score.

#### 3) IMPLEMENTATION DETAILS

The architecture is not too deep, making our model not computationally expensive and resource-intensive. The size of the DAE latent layer equals the number of clusters, which is empirically determined in the first experiment. The ReLU function activates all the internal layers. Finally, the autoencoder is trained using the Stochastic Gradient Descent (SGD) optimizer with a learning rate of 1.0, gradually decreasing, and momentum of 0.9. We set the batch size equal to 50 samples and avoided tuning any hyper-parameters of the model.

### B. COMPARATIVE STUDY

#### 1) BASELINE METHODS

The proposed DJRC is compared with a set of deep and manifold-based clustering methods including state-of-the-art deep clustering methods: k-means [39], deep embedded clustering (DEC) [15], Deep Embedded Regularised Clustering (DEPICT) [34], Deep Adaptive Clustering (DAC) [40], Deep Embedded Dimensionality Reduction Clustering (DERC) [41]. For these methods, the performance results are taken from the original publications.

#### 2) EXPERIMENT RESULTS

Tables 1 and 2 outline the performance in terms of accuracy and NMI, respectively, where the top three accuracy scores are highlighted. Table 1 shows that both variants of DJRC are competitive with other algorithms across all benchmarks. It outperforms the mentioned deep clustering methods. In addition, DEPICT, and DAC techniques are designed for image datasets, so these techniques cannot be performed on other datasets, such as document datasets. In contrast, DIRC can be applied to any dataset.

### C. CASE STUDY USING CORD-19 DATASET

To assess the performance of the proposed method, we conduct experiments on the public CORD-19 dataset. This dataset is made up of scientific documents that are concerned with discussing COVID-19 from different aspects. Note that the databases from which these documents are taken are in different languages, including English and other languages, as shown by Table 3. This Table also mentions the number of documents for each language. As a preliminary step, we detect the language of each document. Then we limit our attention to English documents. Figure 4 represents the number of words in different papers. As can be seen from this figure, most papers are about 5000 words in length. The existence of outliers causes the long tail in Figure 4. Approximately 98% of the papers are under 20,000 words in length.

Following, we report our experimental findings by measuring the performance of our proposed architecture on CORD-19:

- In the first experiment, we evaluate the dimensionality reduction and clustering algorithm in terms of internal validation when varying the number of clusters.
- In the second experiment, we compare the performance of the proposed algorithm against several clustering algorithms.
- The third experiment is an interpretability analysis to shed light on the decision-making process of our model.

**TABLE 1.** DJRC vs. other baselines: accuracy scores.

| Models | Dataset | | |
|---|---|---|---|
| | USPS | MNIST-FULL | MNIST-Test |
| **k-means** | 0.668 | 0.572 | 0.570 |
| **DEC** | 0.619 | 0.843 | 0.841 |
| **DEPICT** | 0.964 | 0.965 | 0.963 |
| **DAC** | 0.972 | 0.978 | 0.975 |
| **DERC** | 0.977 | 0.975 | 0.976 |
| **DJRC** | **0.979** | **0.981** | **0.979** |

**TABLE 2.** DJRC vs. other baselines: NMI scores.

| Models | Dataset | | |
|---|---|---|---|
| | USPS | MNIST-FULL | MNIST-Test |
| **k-means** | 0.450 | 0.499 | 0.498 |
| **DEC** | 0.586 | 0.816 | 0.814 |
| **DEPICT** | 0.927 | 0.917 | 0.915 |
| **DAC** | 0.928 | 0.935 | 0.934 |
| **DERC** | 0.942 | 0.927 | 0.924 |
| **DJRC** | **0.945** | **0.938** | **0.936** |

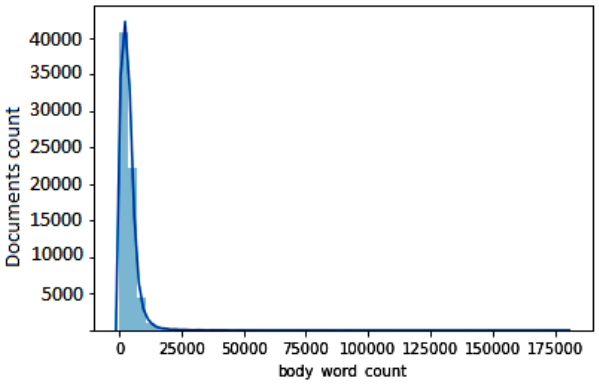



**FIGURE 4.** Count of words in papers.

**TABLE 3.** The number of papers in each language. af: Afrikaans, ca: Catalan, cy: Welsh, de: German, en: English, es: Spanish, fr: French, it: Italian, nl: Dutch, pl: Polish, pt: Portuguese, zh-ch: Simplified Chinese.

| language | af | ca | cy | de | en | es | fr | it | nl | pl | pt | zh-cn |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of papers | 2 | 5 | 7 | 110 | 57602 | 290 | 336 | 18 | 40 | 3 | 15 | 3 |

- In the fourth experiment, we discuss the performance of LDA when varying the number of topics.
- In the last experiment, we visualize the CORD-19 dataset to get a view of the prediction of our model.

### 1) SUITABLE NUMBER OF CLUSTERS

It is very challenging to precisely determine the number of clusters in the CORD-19 dataset. This parameter is quite crucial, as the outcomes of the proposed method are heavily dependent on this parameter. In the case of an unknown number of clusters, the first step is to specify a range of candidates (k) from which the exact number of clusters is picked out. In this work, we set this range to [2, 50]. In this experiment, we evaluate the performance of our algorithm for dimensionality reduction and clustering in terms of DB and SC measures when varying the number of clusters. Figure 5 depicts the performance of the proposed method, in terms of DB and SC metrics, when varying the number of clusters. As shown in Figure 5, the proposed method reaches its best performance for k between 5 and 10. The goal of clustering is to make the distance between two points belonging to two different clusters far apart (Inter-cluster distance) and to make points belonging to the same cluster close as possible (Intra-cluster distance). This is exactly what the Internal validation measures (i.e. DB and SC measures) are targeting to do based on the following two criteria: Separation and Compactness. The Separation measures how distinct or well-separated a cluster is from other clusters. The Compactness measures how closely related the objects in a cluster are. Consequently, the optimal number of clusters is the one for which the proposed algorithm reaches its best performance in DB and SC measures.

### 2) EVALUATING THE PERFORMANCE OF CLUSTERING ALGORITHM

In this experiment, we compare our method with the following methods:

- Deep Embedded Clustering DEC [15]
- Birch [42]
- Spectral Clustering [43]
- COVID-19 LC [12]
- and the one adopted in [14] we called DAE+UMAP+AGG

Our evaluation aims to compare our model to the different approaches to assess the clustering effectiveness, stability, and reliability of our model using the DB and SC metrics. The comparison results are given in Table 4. We can notice through this table that our algorithm is significantly better than all the other algorithms in terms of all indices. These

**TABLE 4.** Comparison with different algorithms in terms of Davies-Bouldin (DB) and Silhouette Coefficient (SC).

| Clustering Methods | Clustering measures | |
|---|---|---|
| | DB | SC |
| COVID-19 LC [12] | 5.83 | 0.016 |
| Birch [42] | 6.50 | 0.007 |
| Spectral clustering [43] | 5.75 | 0.014 |
| DEC [15] | 25.99 | 0.009 |
| DAE+UMAP+AGG [14] | 1.22 | 0.359 |
| Ours | **0.637** | **0.520** |

results assess the quality of our algorithm's clustering, which ensures outstanding dataset organization. We can observe that DAE+UMAP+AGG [14] came in second compared to the other methods in terms of all indices. In DAE+UMAP+AGG, the combination of deep and manifold embedding techniques improves the performance of agglomerative clustering. In DAE+UMAP+AGG, and compared to the other models, the combination of deep and manifold embedding techniques has improved the performance of agglomerative clustering. Although DEC is a deep clustering method that is designed to deal with high-dimensional and non-linear data, we notice that the classical clustering approaches (i.e. Spectral clustering [43], COVID-19 LC [12] which use k-means as a clustering algorithm, and Birch [42]) generally outperform the DEC [15].

The learning process of the proposed deep learning model is based on two kinds of losses, namely reconstruction, and clustering. We can understand how our deep learning algorithm represents the CORD-19 dataset through loss functions. If the predicted output of our model deviates too much from the actual data or output, the loss function will produce a high error value. Figure 6 shows the behavior of the two losses. This figure shows that the two losses softly decreased toward zero. This can be considered a good indication of the performance of our model.

### 3) MODEL INTERPRETABILITY

An important point we should study is understanding the model's performance and interpreting its results. In this part, we use each cluster's feature importance analysis technique. We used a model-agnostic approach called the Unsupervised to Supervised technique [44]. This technique converts the unsupervised clustering problem into a One-vs-All supervised classification problem using an interpretable classifier such as a tree-based model. The steps to do this are as follows:

- Change the cluster labels into One-vs-All binary labels for each
- Train a classifier to discriminate between each cluster and all other clusters
- Extract the feature importance from the model

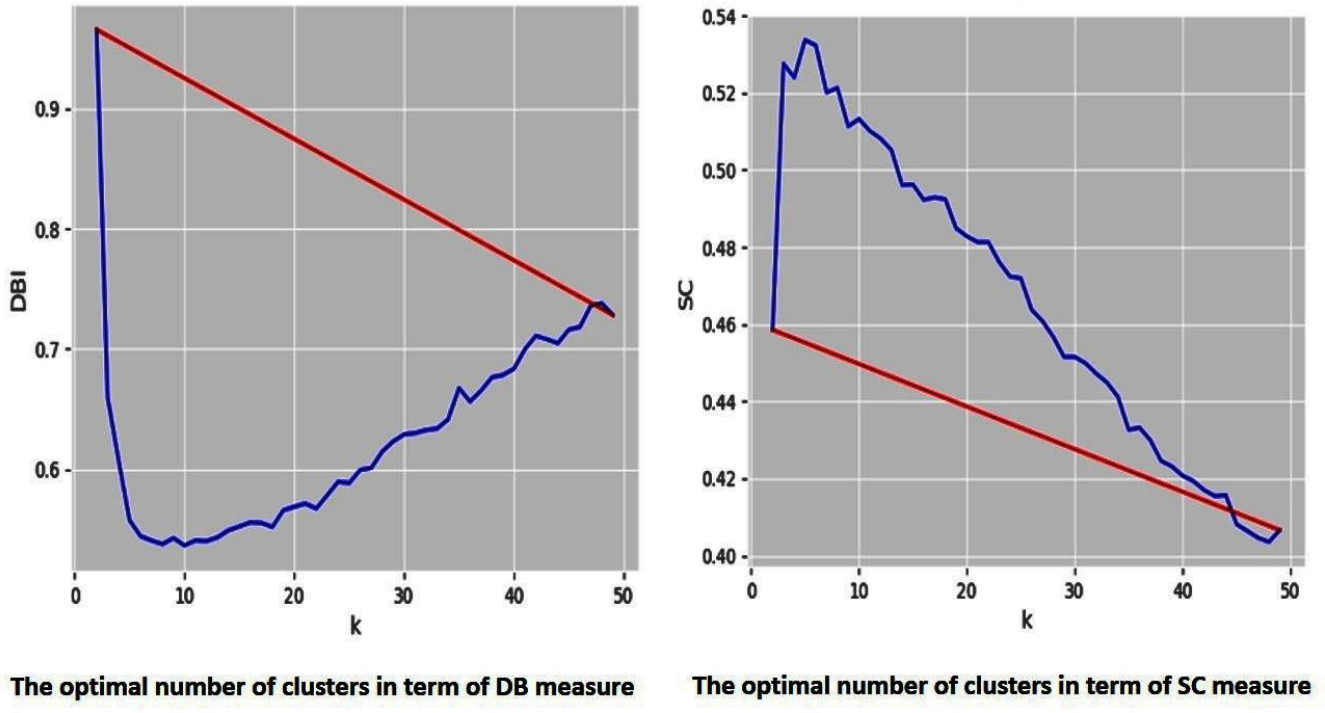



**FIGURE 5.** Evaluating the performance of the proposed algorithm in terms of DB and SC when varying the number of clusters.

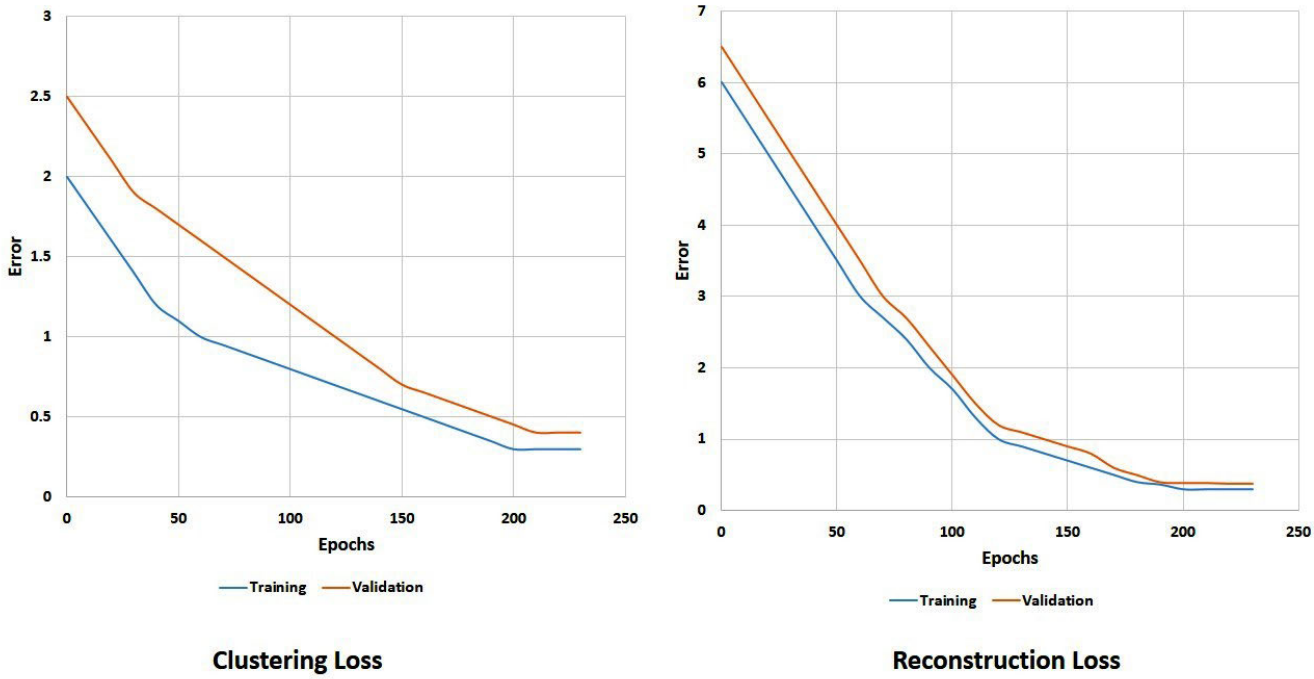



**FIGURE 6.** The error of the clustering and reconstruction loss per epochs.

After converting the problem into a binary classification problem, we chose Random Forest Classifier for the next step, which is the importance of getting the features with the most discriminatory power between all clusters and the targeted cluster. Figures 7 and 8 present the achieved results.

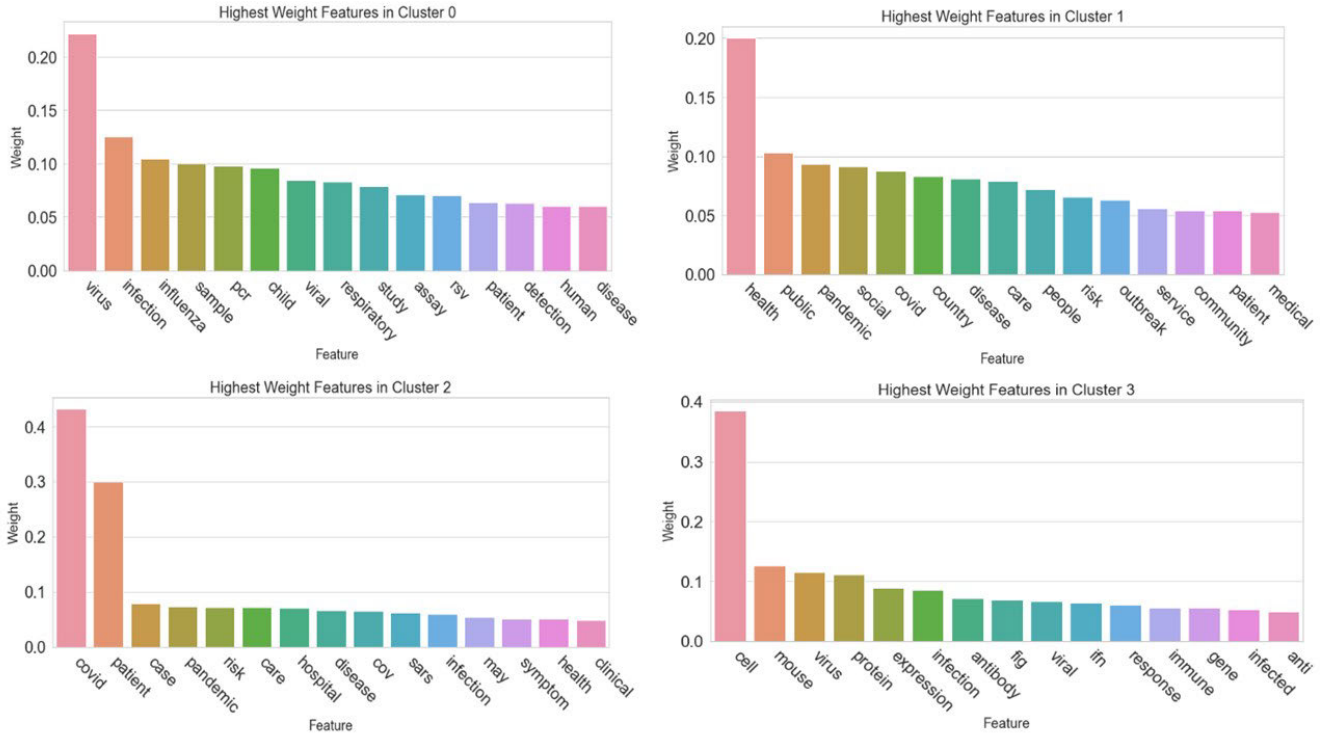



**FIGURE 7.** Most important features using Unsupervised to Supervised method for the clusters 0, 1, 2, and 3.

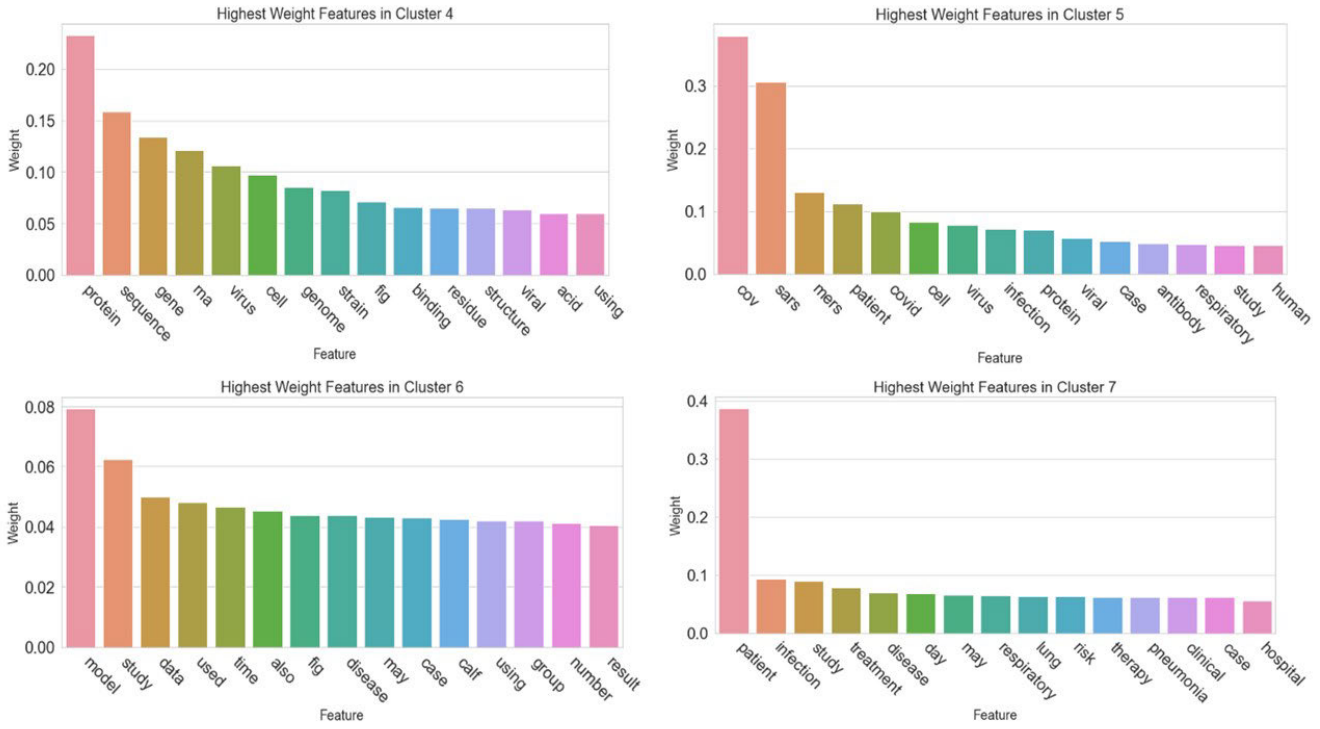



**FIGURE 8.** Most important features using Unsupervised to Supervised method for the clusters 4, 5, 6, and 7.

The significance of Clustering Interpretability becomes evident in scenarios where ground truth labels are absent during the development phase. This absence not only hinders data scientists from directly assessing the validity of clustering using internal validation indices but also complicates the task of conveying cluster performance to stakeholders

in a straightforward and understandable manner. In this context, we have introduced a potential method aimed at addressing this challenge. This method revolves around extracting feature importance specific to clusters, enabling us to comprehend the rationale behind the configuration of each cluster by model. This approach extends to effective communication with stakeholders and intuitive evaluation and finds applications in cluster-based Keyword Extraction within Natural Language Processing (NLP) and as a general technique for feature selection.

#### 4) TOPIC MODELING EVALUATION

In this experiment, we pass the outputs of our deep architecture (i.e., clusters assignment) to LDA, extracting the topics in each document. The aim of this experiment is to find the topics contained in the CORD-19 dataset, which allows us to know if there are biases in the collected data. Table 5 presents the topics predicted by the proposed method for each cluster. Those topics are predicted through the terms that describe each topic. In addition, Figure 9 shows the number of documents that has or share the same topic.

The selected topics were labeled with the names of sub-fields related to Medicine, Biology, and Chemistry. Figure 9 represents the number of labeled topics: Virology, Immunology, Surgery, the Risk factor of COVID-19, Intensive Care Medicine, Molecular Biology, Pathology, and Genetics. It can be seen that the most significant number of publications is in Virology. The relationship of the sub-field of virology to COVID-19 explains this large number of publications, and this is because virology deals with the study of the COVID-19 virus, its variants, and attempts to find the appropriate vaccines. The fields of Immunology, Surgery, and Risk factor of COVID-19 have roughly equal proportions of publications. These fields are considered to be hot topics that have attracted significant attention. These fields studied the behavior and consequences of this virus on health and its impact on citizens, education, the economy, and many other areas. Finally, The least active fields during the COVID-19 pandemic are Molecular Biology, Pathology, and Genetics. The reason for the low number of publications in these areas is due to that these areas are not closely related to the COVID-19 pandemic.

We assess the performance of the LDA when feeding it with the entire dataset (instead of clustered documents) using the Coherence Score $C_V$. The aim was to measure if LDA provides a meaningful, accurate, and latent topic representation. A set of statements or facts is considered coherent if they support each other. Topic Coherence measures a single topic's score by measuring the degree of semantic similarity between high-scoring words in the topic. These measurements help distinguish between semantically interpretable topics and topics that are artifacts of statistical inference. Figure 10 outlines the performance of LDA measured by $C_v$ score while varying the number of topics.

The coherence score seems to keep increasing with the number of topics. In particular, the best performance is reached when the number of topics is equal to 8, which is analogous to the number of clusters detected by our proposed method. The difference, however, lies in the computational cost required by the two strategies. In the proposed method, as we fed the LDA with clustered documents, the computational cost will be much less than fed the whole dataset.

**TABLE 5.** The descriptive terms of each topic per cluster.

| Topic per Cluster | Descriptive terms |
|---|---|
| **Virology** | coronavirus, rabies, response, gene, roost, viral, genome, sars-cov, annotation, bent-winged, herpesvirus, table, epitope, china, supplementary, microbiome, protein, sample, infection, cell, sequence, disease |
| **Immunology** | mouse, viral, antibody, immune, gene, response, culture, day, siRNA, lung, min, target, medium, membrane, patient, autophagy, apoptosis, tumor, expression, cancer, infection, bind, demyelination, tissue, receptor, effect, human, analysis, epithelial, delivery, virus, rat, type, feline, bovine, coronavirus, gut |
| **Surgery** | symptom, health, surgery, china, epidemic, virus, care, treatment, medicine, facemask, endoscopy, procedure, protein, region, humidity, -ent, tube, placement, pack, self-isolate, tip, subcutaneous, case, study, disease, risk |
| **Risk factor of COVID-19** | care, patient, hospital, disaster, outbreak, country, global, human, emergency, information, international, sars, animal, china, population, report, plan, response, research, study, train, ebola, service, infectious, pandemic, medium, nurse, individual, program, case, agent, review, management, biological, education, travel, medicine, food, air, migrant, shortlist, laboratory, surveillance, datum, risk |
| **Intensive care medicine** | medicine, pandemic, patient, hospital, symptom, health, china, epidemic, virus, care, need, woman, pregnant, infection, treatment, medicine, Chinese, facemask, endoscopy, procedure, spike, protein, region, humidity, helmet, absolute, ent, tube, placement, pack, self-isolate, tip, subcutaneous, times, case, study, disease, surgery, datum, risk, use |
| **Molecular biology** | patient, cell, antibody, lung, vaccine, case, day, mouse, hospital, treatment, drug, bind, pro, structure, activity, resistance, genotype, head, polymorphism, allele, animal, sars-cov, human, protein |
| **Pathology** | lung, virus, test, influenza, cause, antibiotic, fever, pneumonia, recipient, pulmonary, level, transplant, pathogen, exacerbation, blood, mortality, health, care, asthma, symptom, gene, bacterial, tissue, infant, child, airway, cell, lesion, signify, sequence, treatment, cancer, transport, centre, wait, group, day, hospital, room |
| **Genetics** | protein, illness, hcov-, nsp, hcov-oc, cns, mouse, disorder, codon, lipid, transmission, ccov, dog, subgroup, cell, hcov, sequence, hcov-nl, genotype, child, patient, gene, strain |

#### 5) VISUALIZATION

To get a better view of the clusters' structure and how our model organized the documents, the CORD-19 dataset was projected into a 2D space using UMAP.

Through Figure 11, It is noticeable that the virology group is widespread and overlaps with all clusters, which can be explained by the relationship of the virology field with all other fields. In addition, some points appear to be outliers, however, they are clustered in the virology field, which can be explained by how similar these points are to the points of this cluster. It can be seen that the clusters of immunology,

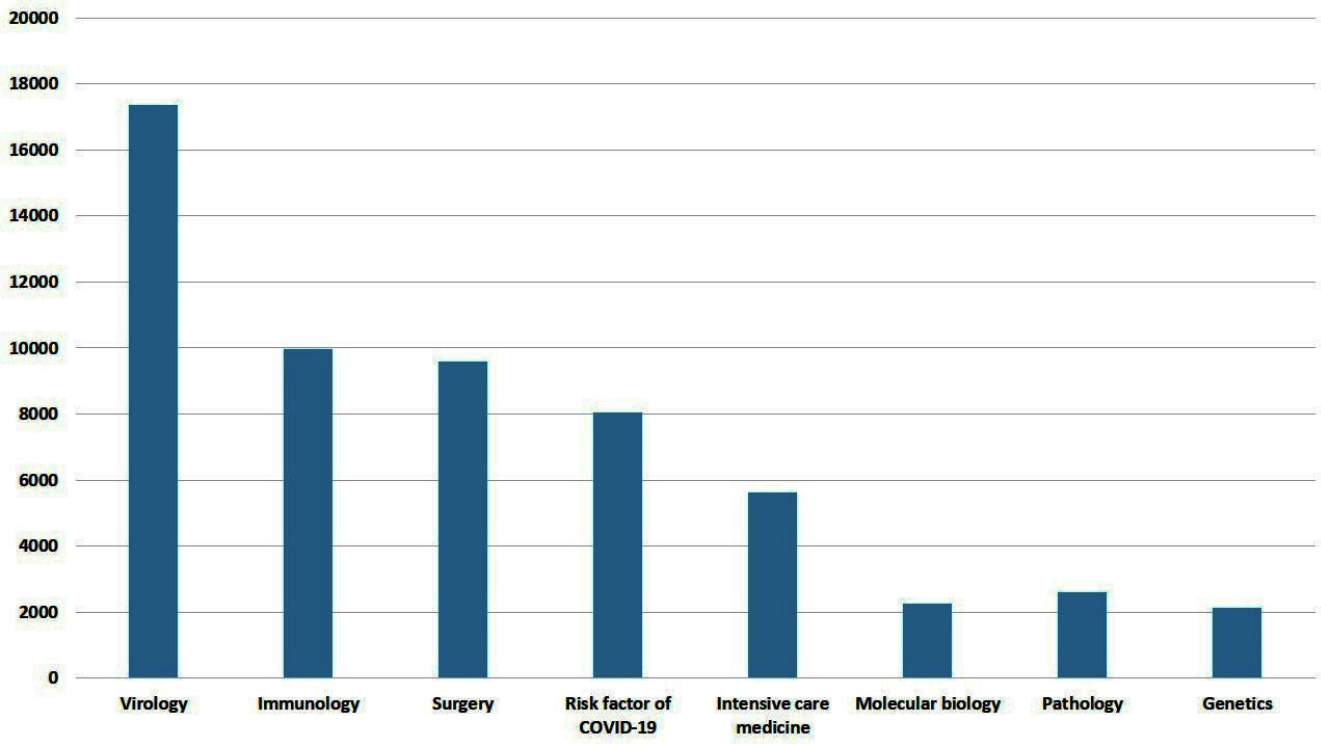



**FIGURE 9.** The number of documents shared the same topics.

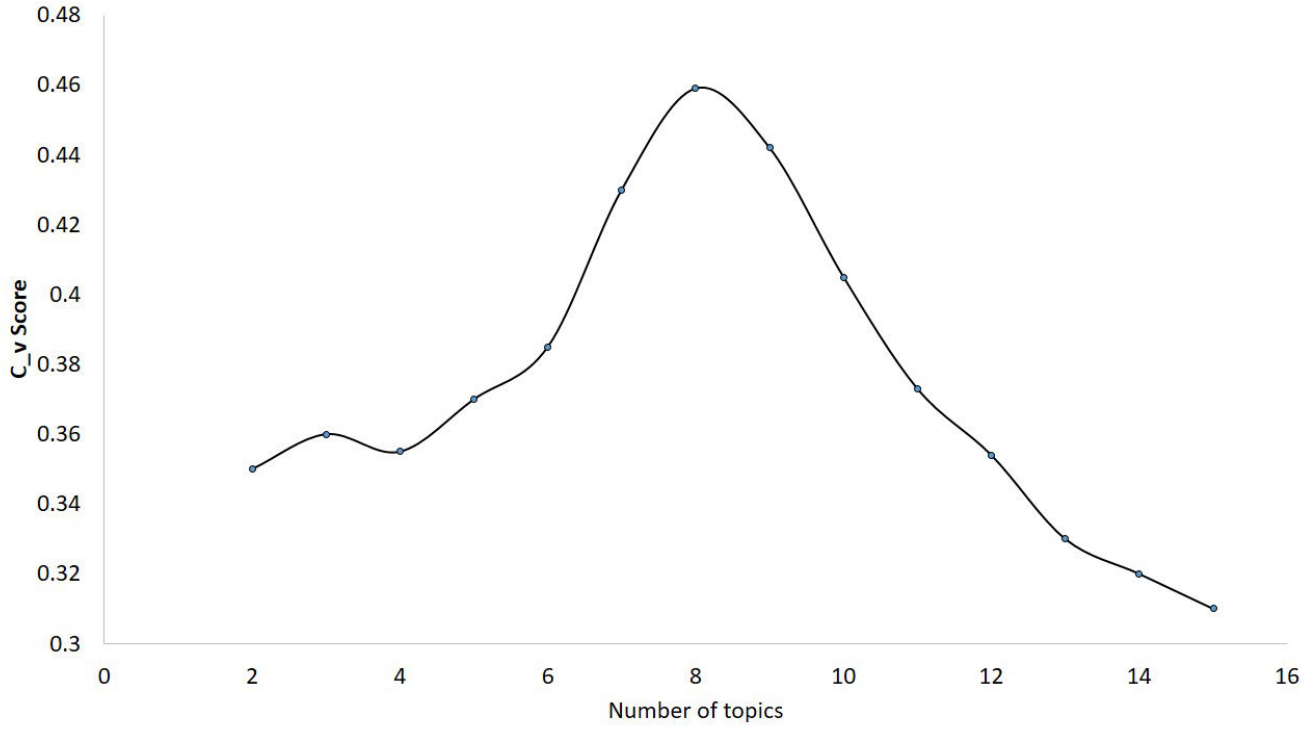



**FIGURE 10.** Measuring the performance of LDA using $C_v$ score while varying the number of topics.

genetics, molecular biology, and pathology overlap, and this is due to the similarity of these fields. The groups for surgery and the risk factor of COVID-19 are separated rather well because these two fields have unique and different keywords from the rest clusters. Ultimately, it is worth mentioning that UMAP is also a dimension reduction method, and its

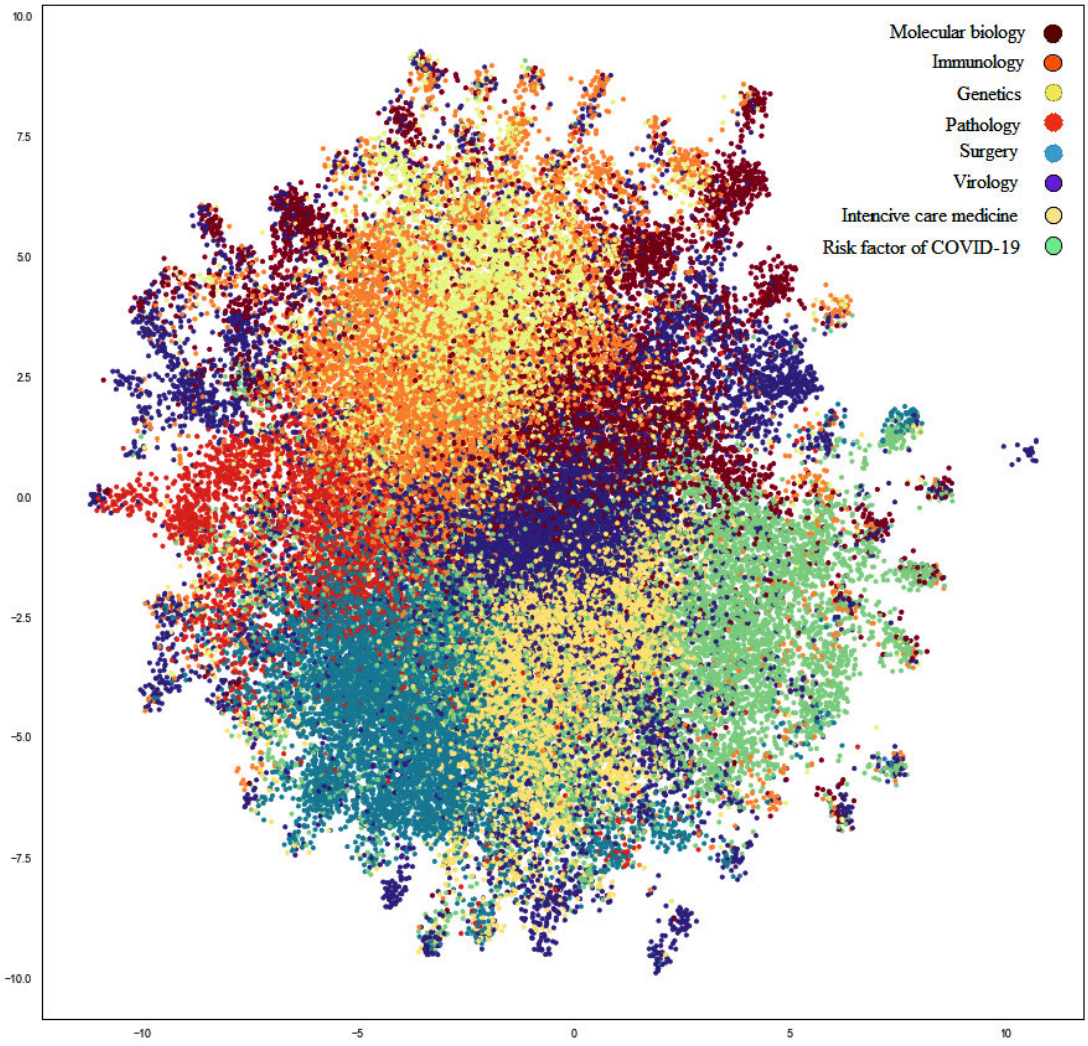



**FIGURE 11.** Visualization of CORD-19 dataset using UMAP.

way of work can affect the latent space produced by our model.

## V. CONCLUSION

This paper proposed a novel deep-learning architecture for organizing a large dataset of COVID-19-related scientific literature. The architecture is made up of three main components, namely two encoders that are jointly trained with one decoder. The main idea behind our model is to train the autoencoder using a two-fold objective function that incorporates two different terms. The first term is the reconstruction loss, designed to check the latent representation, whereas the second term (clustering loss) is dedicated to clustering the input documents. Then, the Latent Dirichlet Allocation (LDA) is used to analyze the document's topics. To improve the computational efficiency of our method, we have considered feeding the LDA with the clustered documents instead of feeding the whole dataset. We conduct thorough experiments on a public dataset. Experimental results show that the proposed method can produce accurate predictions of topics. In addition, experimental results demonstrate that our method has significantly outperformed several recent studies. Although the LDA has produced promising results, as a future direction, one can investigate the possibility of developing a new topic modeling technique based on the recent advances in Natural language processing.

## ACKNOWLEDGMENT

The authors would like to thank Qatar National Library (QNL), for supporting them in publishing their research.

## REFERENCES

[1] *Coronavirus Disease (COVID-2019) R & D*. Accessed: Sep. 2, 2020. [Online]. Available: https://www.who.int/teams/blueprint/covid-19

[2] *COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)*, Coronavirus Resource Center, John Hopkins Univ. & Med., Baltimore, MD, USA, 2020.

[3] *Pneumonia Unknown Cause—China*. Accessed: Sep. 2, 2020. [Online]. Available: https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/

[4] M. S. Diamond and T. C. Pierson, "The challenges of vaccine development against a new virus during a pandemic," *Cell Host Microbe*, vol. 27, no. 5, pp. 699–703, May 2020.

[5] M. Nicola, Z. Alsafi, C. Sohrabi, A. Kerwan, A. Al-Jabir, C. Iosifidis, M. Agha, and R. Agha, "The socio-economic implications of the coronavirus pandemic (COVID-19): A review," *Int. J. Surg.*, vol. 78, pp. 185–193, Jun. 2020.

[6] O. Aiadi and B. Khaldi, "A fast lightweight network for the discrimination of COVID-19 and pulmonary diseases," *Biomed. Signal Process. Control*, vol. 78, Sep. 2022, Art. no. 103925.

[7] M. Jamshidi, A. Lalbakhsh, J. Talla, Z. Peroutka, F. Hadjilooei, P. Lalbakhsh, M. Jamshidi, L. L. Spada, M. Mirmozafari, M. Dehghani, A. Sabet, S. Roshani, S. Roshani, N. Bayat-Makou, B. Mohamadzade, Z. Malek, A. Jamshidi, S. Kiani, H. Hashemi-Dezaki, and W. Mohyuddin, "Artificial intelligence and COVID-19: Deep learning approaches for diagnosis and treatment," *IEEE Access*, vol. 8, pp. 109581–109595, 2020.

[8] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, S. Yang, P. W. Eklund, T. Huynh-The, T. T. Nguyen, Q.-V. Pham, I. Razzak, and E. B. Hsu, "Artificial intelligence in the battle against coronavirus (COVID-19): A survey and future research directions," 2020, *arXiv:2008.07343*.

[9] Q.-V. Pham, D. C. Nguyen, T. Huynh-The, W.-J. Hwang, and P. N. Pathirana, "Artificial intelligence (AI) and big data for coronavirus (COVID-19) pandemic: A survey on the state-of-the-arts," *IEEE Access*, vol. 8, pp. 130820–130839, 2020.

[10] A. Ai, "COVID-19 open research dataset challenge (CORD-19)," Allen Inst. Artif. Intell., Seattle, WA, USA, Tech. Rep., 2020.

[11] B. S. Anderson, "Using text mining to glean insights from COVID-19 literature," *J. Inf. Sci.*, vol. 49, no. 2, 2021, Art. no. 01655515211001661.

[12] M. E. Eren, N. Solovyev, E. Raff, C. Nicholas, and B. Johnson, "COVID-19 Kaggle literature organization," in *Proc. ACM Symp. Document Eng.*, Sep. 2020, pp. 1–4.

[13] S. Reddy, R. Bhaskar, S. Padmanabhan, K. Verspoor, C. Mamillapalli, R. Lahoti, V.-P. Makinen, S. Pradhan, P. Kushwah, and S. Sinha, "Use and validation of text mining and cluster algorithms to derive insights from corona virus disease-2019 (COVID-19) medical literature," *Comput. Methods Programs Biomed. Update*, vol. 1, Jan. 2021, Art. no. 100010.

[14] M. Allaoui, N. E.-H. S. B. Aissa, A. B. Belghith, and M. L. Kherfi, "A machine learning-based tool for exploring COVID-19 scientific literature," in *Proc. Int. Conf. Recent Adv. Math. Informat. (ICRAMI)*, 2021, pp. 1–7.

[15] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 478–487.

[16] A. Youcefa, M. L. Kherfi, B. Khaldi, and O. Aiadi, "Understanding user intention in image retrieval: Generalization selection using multiple concept hierarchies," *Telkomnika*, vol. 17, no. 5, pp. 2572–2586, 2019.

[17] M. Korichi, M. L. Kherfi, M. Batouche, and K. Bouanane, "Extended Bayesian generalization model for understanding user's intention in semantics based images retrieval," *Multimedia Tools Appl.*, vol. 77, no. 23, pp. 31115–31138, Dec. 2018.

[18] F. Debbagh, M. L. Kherfi, and M. C. Babahenini, "Une solution au problème de l'oubli en recherche d'images par les concepts et les relations sémantiques," in *Proc. La Conférence Internationale sur l'Intelligence Artificielle Technologies l'Inf.*, 2014, p. 7.

[19] C. Kim, V. Zhu, J. Obeid, and L. Lenert, "Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke," *PLoS ONE*, vol. 14, no. 2, Feb. 2019, Art. no. e0212778.

[20] F. Zhang, H. Fleyeh, X. Wang, and M. Lu, "Construction site accident analysis using text mining and natural language processing techniques," *Autom. Construct.*, vol. 99, pp. 238–248, Mar. 2019.

[21] H. Leopold, H. van der Aa, J. Offenberg, and H. A. Reijers, "Using hidden Markov models for the accurate linguistic analysis of process model activity labels," *Inf. Syst.*, vol. 83, pp. 30–39, Jul. 2019.

[22] Y. Al Amrani, M. Lazaar, and K. E. El Kadiri, "Random forest and support vector machine based hybrid approach to sentiment analysis," *Proc. Comput. Sci.*, vol. 127, pp. 511–520, Jan. 2018.

[23] Y. Chen, "Convolutional neural network for sentence classification," M.S. thesis, Univ. Waterloo, Waterloo, ON, Canada, 2015.

[24] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," 2014, *arXiv:1404.2188*.

[25] R. Socher, E. Huang, J. Pennin, C. D. Manning, and A. Ng, "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 1–12.

[26] M. Allaoui, M. L. Kherfi, and A. Cheriet, "Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study," in *Proc. Int. Conf. Image Signal Process.* Cham, Switzerland: Springer, 2020, pp. 317–325.

[27] M. Allaoui, M. L. Kherfi, A. Cheriet, and A. Bouchachia, "Unified embedding and clustering," Tech. Rep., 2021.

[28] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.

[29] N. Mrabah, N. M. Khan, R. Ksantini, and Z. Lachiri, "Deep clustering with a dynamic autoencoder: From reconstruction towards centroids construction," *Neural Netw.*, vol. 130, pp. 206–228, Oct. 2020.

[30] L. L. Wang et al., "CORD-19: The COVID-19 open research dataset," 2020, *arXiv:2004.10706*.

[31] K. Al Sharou, Z. Li, and L. Specia, "Towards a better understanding of noise in natural language processing," in *Proc. Conf. Recent Adv. Natural Lang. Process.-Deep Learn. Natural Lang. Process. Methods Appl.*, 2021, pp. 53–62.

[32] J. J. Palop, L. Mucke, and E. D. Roberson, "Quantifying biomarkers of cognitive dysfunction and neuronal network hyperexcitability in mouse models of Alzheimer's disease: Depletion of calcium-dependent proteins and inhibitory hippocampal remodeling," in *Alzheimer's Disease Frontotemporal Dementia*. Cham, Switzerland: Springer, 2010, pp. 245–262.

[33] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 1–16, 2010.

[34] K. G. Dizaji, A. Herandi, C. Deng, W. Cai, and H. Huang, "Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5747–5756.

[35] J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 550–554, May 1994.

[36] L. Deng, "The MNIST database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, Nov. 2012.

[37] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[38] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

[39] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, Oakland, CA, USA, Nov. 1967, vol. 1, no. 233, pp. 281–297.

[40] J. Chang, L. Wang, G. Meng, S. Xiang, and C. Pan, "Deep adaptive image clustering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5880–5888.

[41] Y. Yan, H. Hao, B. Xu, J. Zhao, and F. Shen, "Image clustering via deep embedded dimensionality reduction and probability-based triplet loss," *IEEE Trans. Image Process.*, vol. 29, pp. 5652–5661, 2020.

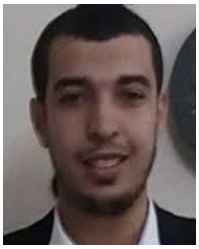
[42] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," *ACM SIGMOD Rec.*, vol. 25, no. 2, pp. 103–114, Jun. 1996.

[43] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 849–856.

[44] O. A. Ismaili, V. Lemaire, and A. Cornuéjols, "A supervised methodology to measure the variables contribution to a clustering," in *Proc. 21st Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2014, pp. 159–166.


**MEBARKA ALLAOUI** (Member, IEEE) received the B.S. and M.Sc. degrees from University Kasdi Merbah Ouargla (UKMO), Algeria, in 2016 and 2018, respectively, where she is currently pursuing the Ph.D. degree in computer vision with the Department of Computer Science and Information Technology. Her current research interests include computer vision and machine learning.

**MOHAMMED LAMINE KHERFI** received the B.S. degree in computer science from the National Institute of Computer Science, Algeria, and the M.Sc. and Ph.D. degrees in computer science from Université de Sherbrooke, Canada. From 2005 to 2013, he was a Professor with the Department of Computer Science, Université du Québec à Trois-Rivières (UQTR), Canada. He was also a Professor with the Department of Computer Science, University Kasdi Merbah, Ouargla, Algeria, and the General Director of digitization and networks with the Ministry of Higher Education. He is currently a Full Professor with the National Higher School of Artificial Intelligence Algeria. He received the U.S. patent. His research interests include image and multimedia processing, computer vision, and machine learning.



**OUSSAMA AIADI** received the M.Sc. and Ph.D. degrees from University Kasdi Merbah Ouargla (UKMO), Algeria, in 2013 and 2017, respectively. He is currently an Associate Professor with the Department of Computer Science and Information Technology, UKMO. His current research interests include computer vision and machine learning.

**SAMIR BRAHIM BELHAOUARI** (Senior Member, IEEE) received the master's degree in telecommunications and network from Institut Nationale Polytechnique of Toulouse, France, in 2000, and the Ph.D. degree in mathematics from the Federal Polytechnic School of Lausanne, Switzerland, in 2006. He is currently an Associate Professor with the Division of Information and Communication Technologies, College of Science and Engineering, Qatar Foundation, Hamad Bin Khalifa University (HBKU). During last years, he also holds several research and teaching positions with Innopolis University, Russia; Alfaisal University, Saudi Arabia; the University of Sharjah, United Arab Emirates; University Technology PETRONAS, Malaysia; and EPFL Federal Swiss School, Switzerland. His research interests include applied mathematics, statistics, and data analysis, artificial intelligence, and image and signal processing (biomedical, bioinformatics, and forecasting). His multidisciplinary background in mathematics and computer science contributes to the comprehensive research endeavors.