Journal of Theoretical and Applied Information Technology
30th November 2020. Vol.98. No 22
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DEEP LEARNING FOR DOCUMENT CLUSTERING: A SURVEY, TAXONOMY AND RESEARCH TREND

LUBNA GABRALLA[1,2], HARUNA CHIROMA[3]

[1] Department of Computer Science and Information Technology, Community College,
Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia,
lagabralla@pnu.edu.sa
[2] Faculty of Mathematical Sciences, University of Khartoum, Sudan
[3] Federal College of Education, computer science department, Gombe, Nigeria

ABSTRACT

Information is stored in several forms such as pictures, web pages, sound and video, but 80% is stored as a text. Quick searching for a specific text document depends totally on the accuracy of the classification of the document's subject with a similar group of documents. This process is called documents clustering. Recently, deep learning techniques have achieved distinguish results in solving the problems facing documents clustering such as complex semantics and high dimensionality. This paper aims to examines a comprehensive review related to documents clustering, and survey the recent work in document clustering using deep learning methods. The proposed taxonomy represents knowledge that helps researchers to understand and follow up previous works in this area, and developing or creating new methods and a comparative analysis was made between popular dataset, performance metrics, deep learning frameworks and library used in deep learning clustering documents.

Keywords: *Clustering, Deep Learning, Documents, Texts*

1. INTRODUCTION

Text data is the most common form compared to other types of data storing on search engines when searching for a specific text document among a collection of large amounts of documents. There are several steps that precede document retrieval, the most important one is document clustering.

Document clustering is a process of organizing documents by dividing them into several separate groups, each group contains documents with similarly related topics and completely different from the other groups. It plays an important role in data mining applications such as knowledge discovery, pattern recognition, and information retrieval.

This process is usually split into two stages: preprocessing and clustering [1]. Preprocessing stage is a critical step to have an easy data. This phase includes cleaning data as the first basic step (also known as Filtering) such as tokenize the document into its component words, remove-stop-words, lemmatization and stemming-word [1, 2], and transform text documents into structured data while the second step is known as feature extraction and feature selection. There are two types of feature extraction methods: the term frequency based method, which is concerned with word count and the semantic web-based method that focuses on words and their relationships [1]. The term frequency–inverse document frequency (TF-IDF) and bag of words (BOW) are the most popular features of extraction models [3]. Feature selection plays a vital role in improving the text clustering performance by reducing the dimensionality and remove redundant information. It is classified into three categories corpus-based method, Latent Semantic Indexing (LSI), and subspace-based clustering [1].

The main objectives of a clustering stage are organizing the data, summarizing it through cluster, groupings in the unlabeled data, and identifying the degree of similarity among points. In literature, more than 100 clustering algorithms have been suggested, and several studies such as [4].[5] [6] [1] classified clustering algorithms to different groups. [7], Clustering methods can be classified into twelve categories, Partitional, Hierarchical, Graph-based, Subspace, Density-based, Constraint based, Groupbased, Neural models, Soft/fuzzy,

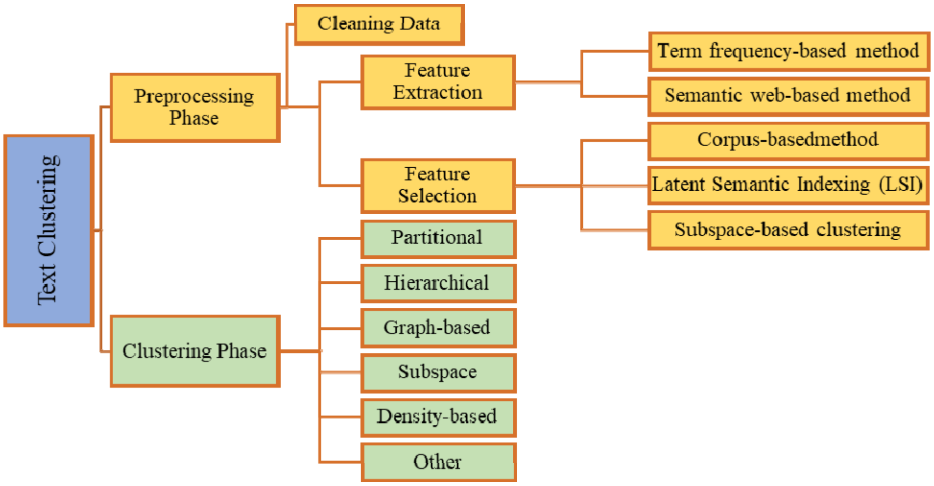

Figure 1: Text Clustering Stages

Distribution-based, Generic approaches and other approaches; however, Partitional, Hierarchical, and Graph-based represent 58% of the total clustering methods used, followed by Subspace (8%) and density-based clustering (7%). Figure 1 displays text clustering stages.

Document clustering is facing several challenges such as complex semantics, big volume, and high dimensionality of the feature space which adversely affects the accuracy of the clustering as some features may be irrelevant [1]. Another issue, the description clustering which is one of the main problems with its traditional algorithm where the result of the clustering cannot give a description of the concept[8]. Deep learning is one of several techniques that have been used to solve this problem and to obtain the most accurate results for document clustering. Current literature pays attention to deep learning clustering (also known as deep embedded clustering or deep clustering). However, previous works mainly focus on reviewing documents clustering or deep clustering independently. That is, this paper included previous studies that have solved a problem of document

clustering using deep learning techniques to help researchers complete the work and achieve the most prominent results in this field.

The main contributions of this work by answered the following research questions:

- ✓ What are the different types of datasets used for document clustering?
- ✓ What are deep learning techniques that used for document clustering?
- ✓ How to measure the quality of documents clusters.
- ✓ What are deep learning performance libraries that used for document clustering?

The rest of this paper is organized as follow: Section 2 examines a comprehensive review related to documents clustering and clustering using deep learning. Section 3 presents fundamental deep learning methods used for clustering documents, subsequently, in Section 4 we propose a taxonomy based on deep learning methods, followed by datasets descriptions in Section 5, then performance metrics is presented in Section 6. Section 7 deals with language, method of converting text to vectors, and domain of applications. Section 8 deals with deep-learning frameworks and library. Section 9 is concerned with the analyses of the results of this study. Challenges and future research direction are in Section10,and conclusion in Section 11.

2. PREVIOUS REVIEW

A considerable amount of literature has been published on survey document clusters. These studies focused on classifying articles according to various clustering techniques. Sathiyakumari et al.[4] provided four clustering categories in document clustering namely Partitional Clustering, Hierarchical Clustering Algorithms, K-Means and Expectation Maximization. Nagma et al.[5] presented a survey on document clustering based on four types on semantic approaches. Similarly

Fahad and Yafooz [9] discussed advantage and disadvantage of several clustering methods in semantic document clustering. Gupta et al.[10] provided systematic review for document clustering based on semantic and traditional approaches. Jensi and Jiji **[11]** reviewed papers of document clustering which focused on soft computing techniques particularly optimization techniques. Mugunthadevi et al [12] presented articles related to feature selection in document clustering. Aggarwal and Zhai [13] provided a broad overview in text clustering specifically in the context of social network. In 2010 Xiao [14] explored summary of document clustering techniques and compared between the mixture of Von Mises-Fisher and Latent Dirichlet Allocation. Patil, and Thakur [15]summarized variety of document clustering techniques used by researchers. On the other hand, there is a relatively small body of literature that is concerned with clustering using deep learning. Aljalbout et al [16] proposed an overview of deep clustering, however, the authors focused on autoencoder methods, and does not take into account other important methodologies. Min et al [17] showed advantages and disadvantages of various deep clustering algorithms via systematic survey of clustering with deep learning based on network architecture. Károly et al.[18] provided explanation of the unsupervised clustering techniques that can be leveraged in deep learning applications. we present similar reviews that have been presented in the literature in this domain based on different clustering approach, deep learning algorithms, and text data set. However, all the studies reviewed so far, did not address articles used deep learning in documents clustering, despite the importance of deep learning that has emerged in various fields, particularly, in documents clustering. Table 1 provides a summary of the reviewed papers.

3 THE RUDIMENTS OF DEEP LEARNING

3.1 Convolutional Neural Network

Traditional neural networks are neural networks that contain one vector as input and group of neurons in a series of hidden layers. Each neuron is linked to all neurons in the preceding layer, and neurons in each layer work independently[21]. This connection leads to powerless, particularly, in object recognition and there is no guarantee that adjacent pixels will be included in the learning [22] in addition to overfitting and complex computation. Con- volutional Neural Network (CNN or ConvNet) is one of the most frequently stated to solve high-dimensional problems and designed to reduce the number of parameters in the network [23]. Furthermore, it is the first successful model representing deep learning in structural concept [24].ConvNet is an important com- ponent in the computer vision, and plays a key role in Natural Language Processing (NLP), sentence modeling, semantic parsing, search query retrieval, and other traditional NLP[25]. It is described by three distinct properties, firstly local connectivity where convolution layer works as similar to the principal of human retina by dividing the image into small parts linked to weights and using this block to help in extracting associated features; secondly, par- ameter sharing convolutional kernels are used to filter the whole image to generate features maps; thirdly, pooling/ sub-sampling which applies non-linear functions to reduce the number of weights and computing linear combination which is the most popular function in max pooling[26]. It returns the maximum value from the slice of the image after using filter. The network is trained and passes gradients through the convolution and pooling. For each feature map, apply discrete convolu- tional operation corresponding to:

$$y_j = g_j \tanh(\Sigma_i k_{ij} * x_i) \qquad (1)$$

Where: x_i is the i^{th} channel of input, k_{ij} is the convolution kernel, and g_j is learned scaling factor. Researchers used several type of activation functions to increase the non-linearity in the image such as tanh, sigmoid, and, Rectified Linear Unit (ReLU). ReLU is distinguished from others by solving the vanishing gradient problem [26]. It is defined by $y = \max(0, x)$The out- put layer is a regular fully connected layer used as the last step in the CNN network to make a non-linear combination of selected features, which are used for the classification of data. ConvNet exchanges between the convolutional layer and pooling layer as illustrated in Figure2

Table 1 summary of the reviewed papers

Reference	Algorithms Covers	Remarks
[4]	Partitional Clustering(k-way)	Divided Clustering algorithms into two categories: Hierarchical algorithms and Partition algorithms only.
	Hierarchical	
	K-Means	
	Expectation Maximization	
[13]	Hierarchical clustering(Single Linkage Clustering, Group-Average Linkage Clustering, Complete Linkage Clustering,	Provide an overview of the traditional clustering algorithms which are often used in the area
	Partitioning algorithms(k-medoid, k-means,	
	Hybrid Approach:Scatter-Gather	
[19]	Autoencoder(Autoencoder ,sparse, stacked , contractive, denoising, variational, graph, convolutional autoencoder, adversarial, and residual)	The authors compared deep learning methods using several datasets
	Deconvolutional Network	
	Restricted Boltzmann Machines(RBM)	
	Deep Belief Nets(DBN)	
[16]	Multilayer Perceptron (MLP)	The study present taxonomy of clustering methods that utilize deep neural networks based on Architecture of main neural network
	Convolutional Neural Network (CNN)	
	Deep Belief Network (DBN)	
	Generative Adversarial Network (GAN)	
	Variational Autoencoder (VAE)	
[17]	Clustering DNN(DNC, DEC, DBC, CCNN, IMSAT, JULE, DAC,	The study focus on clustering with deep learning from the perspective of network architecture
	Autoencoder (DCN, DEN, DSC-Nets, DMC, DEPICT, DCC)	
	Variational Autoencoder-based deep clustering (VaDE, GMVAE)	
	Generative Adversarial Network (DAC, CatGAN, InfoGAN)	
[20]	Partioning-based(K-means, PAM,CLARA, CLARANS, FCM, Discriminative,K-means, DTL-FSSC)	Study provide analytic survey for clustering methods , in addition to descriptive summery for other datamining tasks classification and association rule
	Grid based(WaveCluster, STING, CLIQUE,MAFIA, OptiGrid, O-Cluster)	
	Model based(EM, COBWEB,SOM, CLASSIT)	
	Density based(DBSCAN,OPTICS,DENCLUE)	
	Hierarchial based(AGNES, BIRCH, CURE, ROCK)	
	Fuzzy based(FCM, Subtractive)	
[8]	Fuzzy c-means and Naive Bayes classifier	Comparison between algorithms, tools, and evaluation methods that used for cluster documents based on semantic similarity.
	K-Mean	
	Bisecting k-mean	
	Near Neighbor Clustering technique	
	HAC	
	Fuzzy C-mean with PSO and K-mean with PSO	

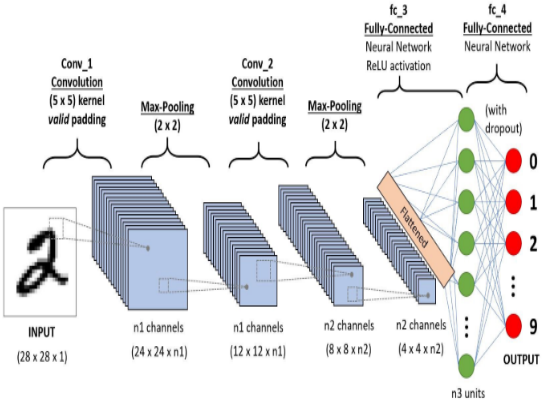

Figure 2: Basic architecture of CNN [27]

There are several architectures of CNNs available, Khan et al[26] categorized deep CNN architectures into seven different classes based on spatial exploitation such as LeNet ,AlexNet, ZefNet ,VGG ,and GoogleNet the main objective of these CNNs is to enhance the performance by optimizing parameters. Depth includes a variety of CNNs to increase performance based on depth of the network, for instance, ResNet, Inception-V3, V4 and Inception-ResNet, ResNex, and Highway Networks. Multi-path proposed to overcome some problems associated with deep networks such as explosion problems or gradient vanishing DenseNets, ResNet, and Highway Networks are a good example for this network. Width based Multi-Connection CNNs is concerned that width is an essential parameter in identifying values of learning beside depth, some popular networks are WideResNet, Pyramidal Net, Xception, and Inception Family. Feature map exploitation researchers proposed models to improve performance by focusing on feature maps, Squeeze and Excitation Network, and CMPESE Network are examples of this models. Channel boosting such as Channel Boosted CNN using TL is a CNN architecture used to extract discriminating features. Finally, features relevant is the main idea behind Residual Attention Neural Network, Convolutional Block Attention Module, Concurrent Spatial and Channel Excitation Mechanism which is classified under Attention.

3.2 Autoencoder

Autoencoder (AE) is an unsupervised learning algorithm applied for generative modelling, feature retraction, and data compression [28]. It includes two parts, the encoder stage by encoding the input x into a latent-space h (also called bottleneck layer or a hidden layer), and the decoder stage reconstructing $r = g_\theta(h)$ instead of predicting the target value in order to copy the output, that means r is expected to be as similar to x as much as possible [17, 29, 30]. Mean square error is used to evaluate the model which can be defined as follow:

$$\min_{\emptyset,\theta} L_{rec} = \min \frac{1}{n}\sum_{i=1}^{n} \left\| x_i - g_\theta\left(f_\emptyset(x_i)\right)\right\|^2 . \quad (2)$$

Where: L_{rec} is reconstruction loss, x is the input vector, $f_\emptyset$ is the encoder part and g_θ is the decoder part. Several methods exist to improve autoencoders ability to gain valuable information and learn the best way for representations such as convolutional autoencoder (CAE), Sparse autoencoder (SAE), and Denoising autoencoder (DAE) .CAE applied convolutional and pooling layers instead of multiple layer perceptions [31].The main aim of CAE is getting features to rebuild the input by learning filters. SAE achieved good results by limiting the number of hidden units or in case the number of hidden units is greater than the number of input, only a small number of the hidden units are allowed to be active at once [30]. Neuron is called active when its output is close to 1 and inactive neuron when its output is to 0 [32].DAE attempts to address identity-function risk by adding noise then autoencoder reconstruct or denoise[17]. A clear indicator of a robust model in DAE is the ability of the model to extract inputs that similar to outputs despite the presence of noises. Figure 3 explains the simple structure of DAE where clean input x yielding $\tilde{x}$ corrupted input ,then $\tilde{x}$ mapped to hidden $y = f_\theta(\tilde{x})$ after that reconstruct $z = g_\theta(y)$ finally $L_H(x, z)$ to minimize the error[33].

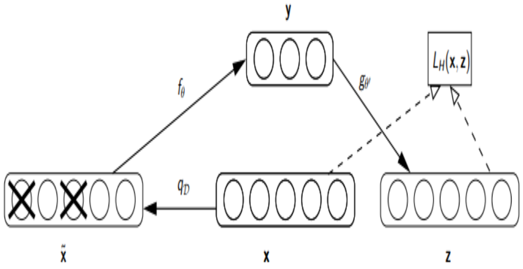

Figure 3: Shows denoising autoencoder [33].

3.3 Deep Belief Networks

Deep Belief Networks (DBN)[34] are probabilistic generative models, designed as multiple layers of Restricted Boltzmann Machine (RBM) which learn based on a probability distribution via its inputs, a layer of the visible unit represents the inputs then follow up by multiple layers of hidden units. The first RBM is trained to reconstruct its input greedily, and from the first hidden layer, we obtain the specific activation probabilities $p(v|h)$ [35] .The hidden layer is considered as a visible layer of the second layer and the second RBM trained using outputs from the previous layer, A learning feature is created as a result of the alternating sampling process between $p(v|h)$, and $p(h|v)$ [35].The last step is adjusting all related weights according to the following equation: Update

$$W_{ij} = W_{ij} + l * \left(P(H_{ij} = 1|V) - P(V_i = 1|H_i)\right) \quad (3)$$

Where l is learning rate, this procedure is repeated for all layers in the network. Figure 4 shows general structure of DBN. It is used as a solution of gradient problem with backpropagation.

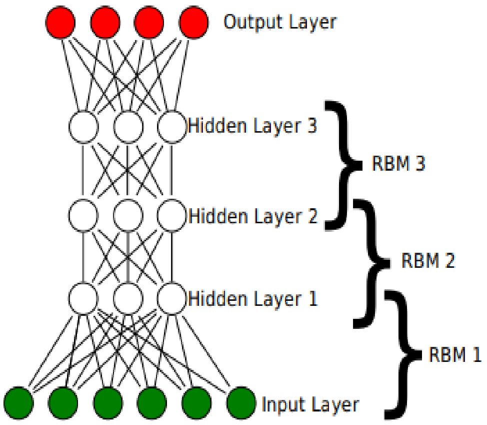

Figure 4: General structure of DBN[36]

3.4 Recurrent Neural Network

Recurrent Neural Network (RNN) is a neural network designed to predict a vector at some time steps by using sequential data[37] .That is, one input sequence vector x_t at time t passes through hidden state h_t to predict output vector y_t at every time step using recurrence formula

$$h_t - f_W(h_{t-1}, x_t) \quad (4)$$

$$h_t = \tanh(W_{hh} h_{t-1} + W_{xh} x_t) \quad (5)$$

$$y_{t=} W_{hy} h_t \quad (6)$$

Where: h_t is new hidden state, and h_{t-1} . RNN is utilized in a variety of applications such as Semantic role labeling (SRL)[38], language modeling[39],and speech recognition[40].

Long Short Term Memory networks (LSTMs)[41] are special kinds of RNN to overcoming vanishing gradient problem by effective track long term dependences and controlling what information passed[42] . LSTMs has four neural network layers in the repeating unit that act in a unique way based on cell state as shown in Figure 5 below.

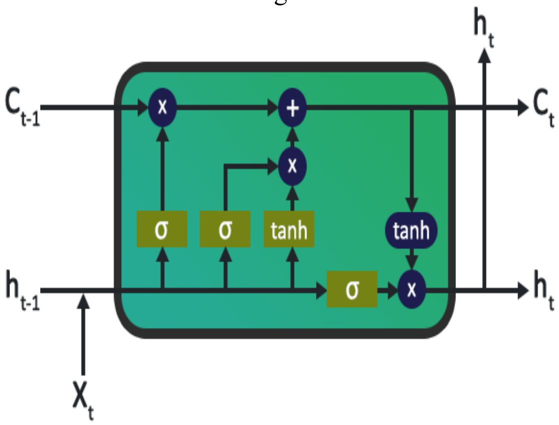

Figure 5: The structure of sequential processing in LSTMs[43]

LSTM deals with cell state by adding or deleting information through sigmoid neural net layer called gates. These gates allow the information pass the outputs in cell state Ct−1 made by sigmoid layer, if output equal zero that means let nothing pass, while one means let everything pass[44]

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f) \quad (7)$$

Two steps to determine storing new data; firstly, update values determined by a sigmoid layer

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i) \quad (8)$$

Secondly, $\tilde{C}_t$new values produced by a tanh layer and added to state

$$\tilde{C}_t = \tanh(W_C . [h_{t-1}, x_t] + b_c) \quad (9)$$

Now the step related to update the old cell state correspond to the following equation where $\tilde{C}_t$ is a new candidate value, new cell state C_t

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (10)$$

Finally, choosing specific output from cell state by using sigmoid layer

$$O_t = \sigma(W_O . [h_{t-1}, x_t] + b_O) \quad (11)$$

$$h_t = O_t * \tanh(C_t) \quad (12)$$

There are numerous alternatives of the LSTM) architecture that have been proposed such as adding peephole connections [45], Depth-Gated LSTM[46], and A clockwork RNN [47].

4 TAXONOMY BASED ON DEEP LEARNING METHODS

Numerous techniques for clustering documents, using deep learning have been developed to date, based on these approaches we will explore each proposed algorithm, indicate to other algorithms compared to them, and their results. As presented in Figure 6, there are twelve deep learning approaches used for documents clustering.

4.1 Convolutional neural network

Kampffmeyer et al [48] designed optimized loss functions to train deep neural networks for clustering based on convolutional neural networks namely deep divergence-based clustering (DDC) and compared the proposed algorithm with several comparative algorithms; the results showed that ensemble (DDC-VOTE) achieves significant results and does not require pre-training steps, other studies applied Convolutional neural networks presented in table 2 below.

4.2 Deep Belief Networks

Chen [52] combined a deep belief network (DBN) for learned features with nonparametric clustering (NMMC). For clustering, the author used the 20 newsgroup dataset for experimental comparison with other baselines. The results proved that (DBN+NMMC) achieved better performance than other methods. Table 3 shows more examples for clustering documents using DBN.

4.3 Recurrent neural network

Frolov et al [54] introduced recurrent neural-network-based Boolean factor analysis for word clustering. The authors used papers published in 2003 and 2004 in the IJCNN, and Neuroinformatics Russian conference in 2004 and 2005, and they applied six different experiments to prove efficiency. Table 4 present articles for documents clustering with recurrent clustering.

4.4 Autoencoder

Current literature pays attention to deep learning clustering (also known as deep embedded clustering) as distinguish methods that merge feature learning and clustering successfully. That is, extracting latent features and cluster original images simultaneously[59]. For instance, Xie, et al[60] applied Deep Embedded Clustering (DEC) using stacked autoencoder (SAE) for parameter initialization then Clustering with KL divergence, to assess DEC they utilized two image datasets and one text dataset in comparison to other algorithms, the results of this study indicate that DEC achieved significant results. However, there is no guarantee that the clustering process included the samples in the margin although clustering loss is the main contribution of DEC[61].Guo et al [61]proposed Improved Deep Embedded Clustering (IDEC) which focuses on data structure preservation by joining clustering loss as direction for scattering data points in feature space and autoencoder. Empirical experiments compared k-means, Spectral Embedded Clustering (SEC), AE+k-means, DEC, and IDEC, the results have proven that IDEC outperformed other methods, and structure preservation plays an important role in the enhancement of deep clustering performance. Table 5 summarizes several Autoencoder proposed techniques.

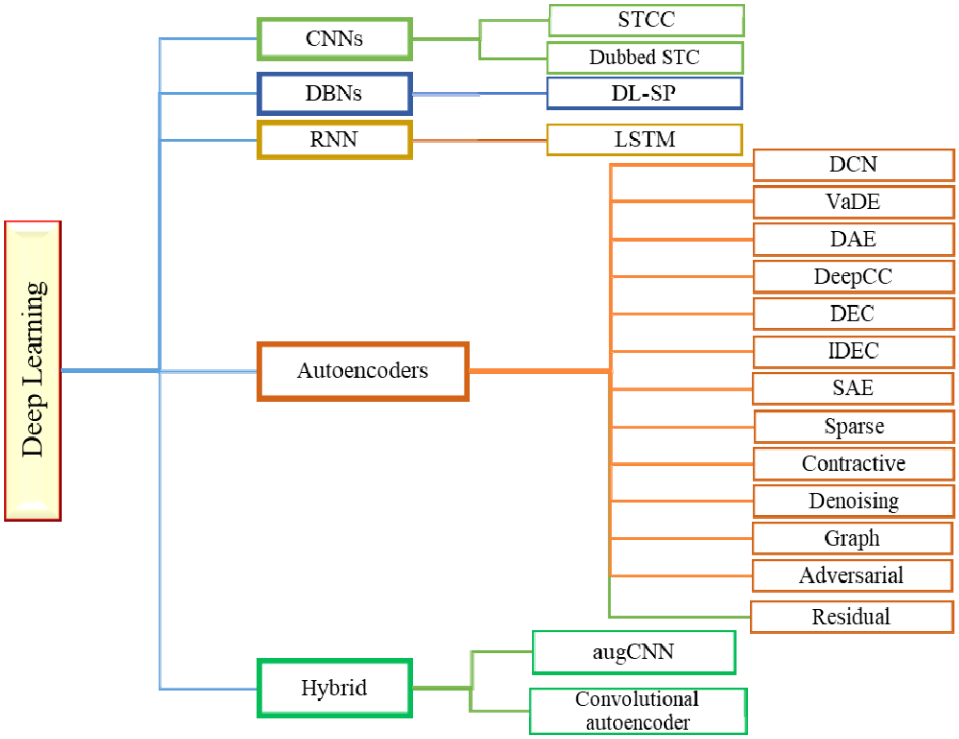

Figure 6 Deep learning approaches used for documents clustering

Table 2 Convolutional Neural Networks for Clustering Text Documents

Reference	Propose deep learning algorithm	Problem	Algorithms compared with	Result
[49]	Short Text Clustering via Convolutional neural networks (STCC)	Short text clustering	- K-means - Spectral Clustering - Average Embedding	STCC achieve significantly better performance than the baseline methods
[50]	Self-Taught Convolutional Neural Networks for Short Text Clustering (dubbed STC2)	Short text clustering	- K-means - SkipVec - RecNN - Para2vec - AE - LSA - LE - LPI - bi-RNN	Dubbed STC2 approach get a significantly better performance than other methods
[51]	Semi Convolutional Neural Networks	Short text clustering	- k-means - LSTM	Semi-CNN more effective than other competitors.

Table3 Clustering Documents using DBN

Reference	Propose deep	Problem	Algorithms compared with	Result

	learning algorithm			
[1]	Sparse group Deep Belief Networks, called Deep-Learning Single-Pass (DL-SP).	Text clustering, feature extraction and data dimension reduction.	- Single-Pass - LSI	DL-SP has better performance than LSI and Single-Pass
[53]	Deep Belief Networks	Clustering Arabic words into split groups	-Naive Bayes(El-Kourdi et al., 2004) - SVM (Hmeidietal.,2008) -kNN (Mohammad AH et al, 2016)	DB more effective than other competitors

Table4 Documents clustering using recurrent neural networks

Reference	Propose deep learning algorithm	Problem	Algorithms compared with	Result
[55]	Recurrent Neural Networks	Measure distributed representations	TF-IDF+ Term-based TF-IDF+ Word-based	TF-IDF with term based distributed concept representation superior other two methods
[56]	Long-term short-term memory network (LSTM)	Pretrain the deep representation model and text clustering	- CNN - BoW - VSM (TF-TDF) - SO - STCC	LSTM superior to other similar algorithms
[57]	Long-term short-term memory network (LSTM)	Documents clustering	- K-means -LDA - GSDMM - GSDPMM	Outperforms other methods
[58]	Recurrent Neural Networks	Extraction semantically meaningful features through clustering models	- k -means - GMM - BIRCH - DEC	Proposed method superior other existing approaches

4.5 Hybrid deep learning algorithms

Numerous studies have assessed the value of combining two or more deep learning techniques, for instance, Guo et al.[59] developed Deep Convolutional Embedded Clustering (DCEC) using Convolutional Autoencoders (CAE) instead of Stacked Autoencoders (SAE). They claimed that SAE unsuccessful for dealing with images while CAE designed for learning features from unlabeled images, researchers used three benchmark datasets and experiments to show that DCEC is more accurate among other methods. Dang et al[70] discovered a non-trivial feature of sentences by using threshold autoencoder and convolutional neural network. The authors compared proposed algorithm with LogReg, MLP and CNN on legal and contract documents. This hybrid algorithms achieved better performance and results. Wang et al [19] compared a variety of deep learning methods including convolutional autoencoder which uses convolutional operations instead of nonlinear activation functions with baseline methods. The authors utilized eight different datasets; the experiments results proved that deep learning methods outperform other methods.

Table 5 Autoencoder proposed techniques

Reference	Propose deep learning algorithm	Problem	Algorithms compared with	Result
[62]	Deep Clustering Network (DCN)	Automatically map high-dimensional data to a latent space	- K-means - SC - SSC-OMP - LCCF - XRAY - NMF+KM - SAE+KM - JNKM - DEC	Effectiveness of DCN
[63]	Variational Deep Embedding (VaDE)	Probabilistic clustering problems	- GMM - AE+GMM - VAE+GMM - LDMGI - AAE - DEC	VaDE outperforms the other methods VaDE's have capability of generating samples for specified cluster, without using supervised information
[64]	Deep Cluster and Gaussian Mixture Model Integrating GMM into DAE (DC-GMM)	Learning representation and clustering analysis.	- K-means - DAEC - GMM - DEC - DAE+Kmeans - DAE+GMM	DC-GMM outperforms the other methods
[65]	Standard stacked autoencoder	Overcomes scalability and generalization of the spectral embedding	- DEC - DCN - VaDE - JULE - DEPICT - IMSAT	SpectralNet outperforms other methods, and is competitive with IMSAT
[66]	Deep autoencoder (DeepCC)	Dimension reduction and clustering	- k-means - SCC - CCInfo	Experimental results demonstrate the effectiveness of DeepCC.
[67]	Deep Embedding Clustering(DEC)	Patent document clustering	-TF-IDF+K-means - TF-IDF + GMM - Bag-of-words + K-means - Bag-of-words + GMM - Doc2Vec + K-means - Doc2Vec + GMM	Doc2Vec+DEC outperforms other methods
[68]	Semi-supervised network embedding (SNE) using stacked auto encoders	Information, text features and category attributes into embedding vectors	-Deep-Walk, Node2ve -GraRep -SDNE -TADW -Text-SVD -SNE-NTC	SNE significantly outperforms
[69]	SIF embedding using autoencoder	Short text clustering	- TF - TF-IDF - Skip-Thought - SIF - STC2	Proposed methods demonstrate the effectiveness
[19]	Autoencoder(sparse, stacked , contractive, denoising, variational, graph, convolutional autoencoder, adversarial, and residual)	Representation for Text Categorization	- Baseline method - Deconvolutional networks - RBM - DBN	Deep learning methods outperforms other methods

5 DOCUMENT DATASET

For deep clustering documents, all algorithms are implemented on a different set of statistical data to judge the effectiveness of utilizing these techniques. In this section, we review these real text data sets in some detail in Table 6.

6 PERFORMANCE METRICS

Numerous evaluation metrics are utilized to evaluate the documents clustering performance, for instance, unsupervised clustering accuracy (ACC) where m refers to mapping function, c means cluster output and y represents truth labels.

$$ACC = max_m \frac{\sum_{i=1}^{n} 1\{y_i = m(c_i)\}}{n} \qquad (13)$$

Another popular performance metrics is normalized mutual information (NMI) is normalized by the average of entropy of both actual labels and the cluster values [50]. Table 7 shows the performance metrics used in the literature.

$$NMI - (Y,C) - \frac{I(Y,C)}{\frac{1}{2}[H(Y) + H(C)]} \qquad (14)$$

7 LANGUAGE

This section reviews the three major aspects related to documents clustering namely language of the documents, methods of converting a text to vectors (word embedding), and domain of applications.

✓ Documents using international languages or local languages, international languages (world languages) which is a language spoken by a large number of people around the world. These languages are used in universities, conferences, global meetings, business and political events. It is defined as a global language based on the total number and geographical distribution of speakers [72]. According to United Nations (UN), the top six languages and official languages for UN are English, Chinese, French, Spanish, Arabic, and Russian [73]. Table 8 presents the Language of the documents in the previous studies, according to[74], Koren is also included in the list of top 20 of the most common languages in the world and widely spoken languages in 2009. On the other hand, local languages (also known regional language) are languages spoken in an area of a nation- state, whether it is a small area, a federal state, an administrative division, or some wider area [74] such as Hindi and Persian. Most of the recent text categorization techniques focused on English and Chinese to validate their methods whereas works on Arabic, Urdu, Hebrew, and Persian are extremely unique in addition to significant differences in dialects among these languages makes authenticating techniques for popular Middle Eastern languages difficult [53].

✓ Distributed representations of words (also known as word embedding) plays a vital role in extracting the semantic of words then achieving significant improvements in clustering documents. Word embedding can be classified into two types, Frequency based and Context based. TF-IDF is a type of frequency-based word embedding, it works by scanning all the sentences in the corpus and taking into account the common words such as - "a", "the", etc. then giving more weight to the words that occur in a smaller subset of documents; however, this method consumes long time and space, while recent trends in modeling distributed representations, for instance, Word2Vec[75] is a type of Context based; it gives examples of neural network using the linear transformation matrix information (convert text set into vector) these models learn the weights of the hidden layer which is fast to train with possible word embedding. There are several common word embedding methods such as Doc2Vec [76], GloVe[77], and ELMo[78]. Table 9 presents advantages and disadvantages of distributed representations of words algorithms used in documents deep clustering studies.

Table 6 Real text data sets for deep clustering documents

Reference	Source	Data Size	Data Description
[19, 48, 60, 63, 65] [58]	Reuters dataset 43M	810000 English news stories	Manually categorized into a category tree, four root categories using as labels, removed stories that are labeled with multiple root categories and represent each news story as a feature vector consisting of the TF-IDF of the 2000 most frequently occurring word stems
[60, 61, 64] [58]	Reuters10k	Sampled a random subset of 10000 examples from Reuters	A copy is available to researchers from well-known Reuters dataset since some algorithms do not scale to the full Reuters datase
[19, 62]	Reuters Corpus Volume 1 Version 2 (RCV1-v2)	Contains 20 topics and 365, 968 documents and each document have a single topic label. And pick the 2,000 most frequently used words	A copy is available to researchers from Reuters
[19, 52, 56, 62] [68] [57]	20Newsgroup corpus	69MB	Collection of 18,846 text documents which are partitioned into 20 different newsgroups It was originally collected by Ken Lang[71]
[55]	U.S. National Library of Medicine (United States) MEDLINE	Contains 24,358,442 records and contains 122,305,477,309 bytes	U.S. National Library of Medicine (United States) contains journal citations and abstracts for biomedical literature from around the world
[55] [68]	PubMed Central (PMC)	PMC contains more than 5 million full-text records, spanning several centuries of biomedical and life science research (late 1700s to present)	a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM) PMC was developed and is managed by NLM's National Center for Biotechnology Information (NCBI).
[49, 50] [69]	SearchSnippets	12340 number 8 classes	Chosen from the results of web search transaction using predefined phrases of 8 different domains
[49, 50] [69]	StackOverflow	20000 number 20 classes	BigQuery dataset includes an archive of Stack Overflow content, including posts, votes, tags, and badges. This dataset is updated to mirror the Stack Overflow content on the Internet archive and is also available through the Stack Exchange Data Explorer.
[1]	TanCorpV1.0	Not mentioned	Not mentioned
[1]	Encyclopedia of China	Not mentioned	Not mentioned
[1]	Sogou Corpus	Not mentioned	2,909,551 news articles from the SogouCA and SogouCS news corpora, in 5 categories, the Chinese characters have been converted to Pinyin.
[50] [69]	Biomedical	20000 number 20 classes	Large scale online biomedical semantic indexing
[51]	question type	5,953 6 class	the TREC question dataset where all the questions are classified into 6 categories: abbreviation, description, entity, human, location and numeric
[51]	Ag_news	4,000 4 class	contains short texts extracted from the AG's news corpus, where all the texts are classified into 4 categories: World, Sports, Business, and Sci/Tech
[51]	dbpedia	14,000 14 class	is the DBpedia ontology dataset, which is constructed by picking 14 non-overlapping classes from DBpedia 2014
[51] [58]	Yahoo_answer	Number of examples is 10,000 Number of classes is 10	10 topics classification dataset extracted from Yahoo! Answers Comprehensive Questions and Answers version 1.0 dataset

[53]	Al-Jazeera	10,000 documents	AlJazeera news website
[53]	Saudi Press Agency	No of text 1,526	Provides its newswires in six class news Cultural, Sports, Social, Economic, Political, and General which were used to label the texts collected from the SPA official web site
[70]	Software Procurement contracts	Not mentioned	Obtained from a major public enterprise with labeling data were annotated by the contract professionals.
[19]	TOX	Each document is 5748-D feature vector.	A gene database of 171 genetic toxicology text documents with four categories
[54]	IJCNN	M=1042 articles	Articles from 2003 and 2004 International Joint Conference on Neural Networks
[54]	Russian Conference	M=189 articls	Articles from the 2004 and 2005 Russian Conference on Neuroinformatics
[19]	CNAE	857 attributes Names 2.2K Data 1.8M	Containing 1080 documents of free text business descriptions of Brazilian companies categorized into a subset of 9 categories
[66] [68]	Citeseer .	3312 documents and each document is described by 3703 words	Contains 3312 documents over the 6 labels (Agents,IR,DB,AI,HCI,ML). It is made of 4 views (content,inbound,outbound,cites). The documents are described by 3703 words in the content view, and by the 4732 links between them in the inbound, outbound and cites views.
[67]	KIPRIS	Consists of abstracts from five categories	Not mentioned
[67]	KISTA	Contains three technology trend reports from the communications field.	Not mentioned
[68]	Cora	4.6 MB	2708 scientific publications classified into seven classes
[68]	Wiki	2405 documents divided into 19 classes and 17981 edges between them.	Not mentioned
[57]	Google News Title Set (T-Set) .	Dataset contains short documents with an average length less than 10	Not mentioned
[57]	Google News Snippet Set (S-Set)	Not mentioned	Not mentioned
[57]	Tweet Set4	Dataset contains short documents with an average length less than 10	Not mentioned
[58]	The Stanford question answer dataset (SQuAD)	Number of examples is 1094 Number of classes is 10	Dataset for reading comprehension that comprises questions and corresponding answers for reading passages of Wikipedia articles
[58]	FakeNewsAMT	Number of examples is 480 Number of classes is 12	Comprises legitimate news articles belonging to six different domains from a variety of mainstream news websites. Also, it contains fake news collected through crowdsourcing via Amazon Mechanical Turk (AMT) for each corresponding domain.

Table 7. Documents Clustering Performance Metrics

Reference	Accuracy Metrics
[48-51, 56, 60-65] [66] [69] [58] [19]	Unsupervised Clustering Accuracy (ACC)
[48-51, 56, 61, 62, 64, 65] [66] [57] [69] [58] [19]	Normalized Mutual Information (NMI)
[1] [55]	F_measure
[55] [64]	Purity
[62, 64] [52] [67] [19]	Adjusted Rand Index (ARI)
[53, 70]	Precision

Table8: Language

Reference	Language of the document	Category of the language
[55] [70] [54] [48] [19, 60, 63, 65] [49] [50] [51] [52] [56] [62] [66] [57] [69] [58]	English	International
[1]	China	International
[53]	Arabic	International
[54]	Russian	International
[67]	Korean	International

Table 9. Vectorization Methods

Reference	Vectorization method	advantages	disadvantage
[49] [50] [51] [56] [57] [69]	Word2vec	Simple and easy to understand, give fairly good performance and taking into account the context of the words suitable for the data sequence faster and saving storage and computing resources can be applied in a variety of tasks NLP.	As the word and vector is one to one relationship, so polysemous problem cannot be solved.Word2vec is a static way, but unable to do specific tasks for dynamic optimization
[62] [63] [64] [65] [67] [68] [48] [1] [53]	TF-IDF	Simple implementation, strong explanatory and easy to understand algorithm	The position information of the word cannot be reflected The accuracy is not high
[67]	Doc2Vec	Inherit the semantics of word vectors and take word order into account	The quality of the vectors is highly dependent on the quality of the word vectors and information is in the paragraph vectors is unclear and difficult to interpret
[67]	Bag-of-words	Simple to understand and implement and Encodes text rather than word The length of the encoded vector is the length of the dictionary	Fails in word ordering and it cannot distinguish common words
[57]	Continuous bag-of-words(CBOW)	It is low on memory Being probabilistic is nature, may superior to deterministic methods	CBOW training if not properly optimized can take forever. CBOW takes the average of the context of a word
[58]	BERT	Take into account word order Having a fixed size vocabulary being able to load the vectors in a GPU regardless of corpus size, can capture context from all possible directions (fully connected). faster training time	Makes it compute-intensive and hard to bring into production

Table 10. Domain of applications

Reference	Domain
[55] [50] [19] [68]	Medical and Biomedical
[53] [51] [52] [49] [56] [60] [61] [62] [63] [64] [65]	News
[70]	Low Documents
[19] [58]	Business
[67]	Technology

✓ Application Domain Recently, the internet allowed the spread of published articles and research at a rapid rate that exceeded millions of those documents in a different application domain. Clustering researches, articles, and documents into clusters or groups, where each group has similar features plays a prominent role in extracting information in a short time, especially in sensitive scientific fields such as Medical and Biomedical domains where text mining techniques such as search and retrieval of documents, natural language processing, text clustering, and text classification in cancer research, contribute to the diagnosis,treatment, and prevention of cancer [73]. News domain contains articles or documents collected from different sources such as World news, Sports news, Business news, and Science/Tech news. Law domain is concerned about legal industries such as contracts regulatory, documents for review, records of goods sold or service, criminal law and juridical sciences. Business domain is another important field, which focused on knowledge points in business sector such as manufacturing, banking, and insurance that mean the documents in business domain are related to understanding all processes, procedures and other key aspects of economic activities. Technology domain is the most recent and broadest field that includes features of all the modern technical and electronic products such as cars, cameras, computers, mobiles and so on. Table 10 shows popular application domains for documents cluster in the literature.

8 DEEP LEARNING FRAMEWORKS AND LIBRARY

Another significant aspect of documents is deep learning frameworks and library. Scientific computing libraries play a vital role in deep learning performance that is why researchers used wide variety of programming frameworks and library for documents clustering. Table 11 displays them in more details.

9 ANALYSIS AND DISCUSSION

Clustering is an unsupervised machine learning used to identify homogeneous subgroups by setting each object in the same group (clusters). That is, each group has similar features than the other clusters. In reviewing the literature, various terms have emerged such as document clustering[81, 82], text clustering[1], text document clustering[83], short text clustering[50, 84],text information[68],topic clustering [85], text categorization[86] , and word clustering[87]. The main purpose behind all these studies is to conduct a clustering process on texts.

Classification is another popular machine learning which is applied in text documents. The main difference between clustering and classification is that in clustering there are no predefined class labels while classifying the data with class labels, but some studies such as [85] combines between the two techniques by using clustering to find the labels then using these labels to classify.

Deep learning has achieved successful result in several text problems such as documents classification, information retrieval, documents modeling, document representation and documents summarizing. This research focused on surveying documents clustering despite some similarity with other concepts and approaches. The main purpose of this study is to highlight the importance of clustering of documents using deep learning techniques.

According to previous works, Reuters dataset and 20Newsgroup are popular data set used in document clustering while ACC and NMI are used in most studies to evaluate the documents clustering performance from the review conducted, figure 7 and figure 8 present comparisons respectively.

Table 11. Programming frameworks and library

Reference	Framework/library	Description	Strength	Limitation
[60, 64]	Caffe[79]	The framework is Python based API and written in c++ for machine learning and deep learning purpose and good choice for computer vision and image processing	Switch between CPU and GPU, have powers academic community speed Models are configured instead of coded	Poor documentation inappropriate for RNN
[62]	Theano[80]	Written in +Python major use of a computational graph and common in the research community	Open source tools, Parallel execution, Support for both CPU and GPU, Support tensors and support algebra operations	Unclear error messages, longer compile times, Use of single GPU and low level
[59, 61, 63, 67]	Keras	Written in Python and touted as a simple interface to full Deep Learning platforms Theano, TensorFlow, CNTK	Easy to learn, integrates with lower-level deep learning languages, has broad adoption in the industry and the research community[80]	Need extra code for complex network Less projects available online No RBM for example
[56, 67, 69, 70]	Tensorflow [63]	Open source framework used for machine learning applications designed by Google researchers Low-level core (C++/CUDA)	Graph visualizations, library management, pipelining, offers great debugging method and scalability [54]	*1) Not support windows, lacks in speed and usage when compared with other framework, computation speed, no GPU support and missing symbolic loops*
[50, 64]	Matlab	It is a high-level language for analyze,data computation, develop algorithms,visualization, and programming	Flexibility,Documentation Use a large database,Create images and videos easily. Debug easily	Very slow, it is more expensive, not suitable for development activities, converting Matlab code to other language need deep knowledge to deal with errors produced
[60, 61, 63, 65-67]	Python	Python developed by Guido Van Rossum in 1991 it is a high-level, the syntax in Python helps the programmers to do coding in fewer steps as compared to Java or C++[81]	Easy to learn,Powerful libraries such as matplotlib and scikit-learn Open source and free Run immediately, Support for OOP, Less coding required Dynamic	Speed limitation,Weak in mobile computing Design restrictions Application Portability Underdeveloped DB Layers

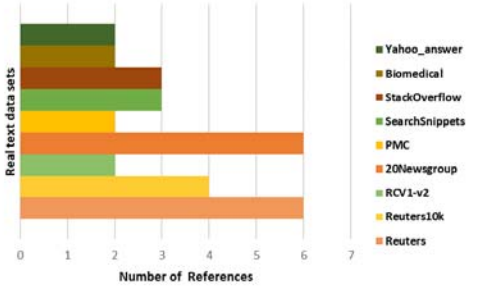

Figure 7: Comparison between most used real text data sets

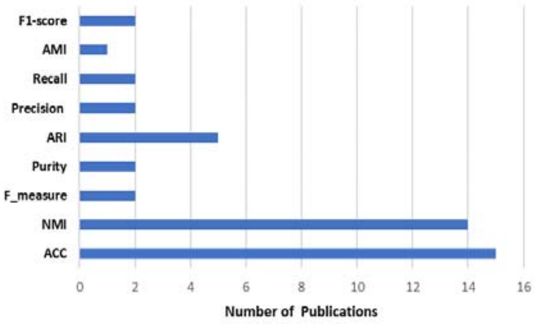

Figure 8: Comparison between documents clustering performance metrics

We can say that research in this field only started in early 2015 with little interest. The field began to attract attention in 2016 with a noticeable increase in 2017, 2018 and 2019. In addition to that, the first quarter of 2020 recorded a reasonable number of studies and is expected to gain significant interest from the research community, and research will increase during the coming years. Figure 9 shows the publication during the period from 2015 to 2020.

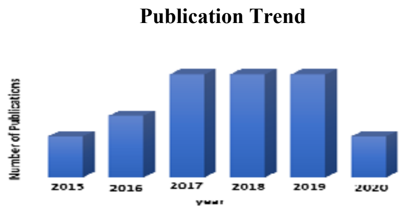

Figure 9: Trend of publications of deep learning in clustering documents

From the perspective of the deep learning methods, pie chart in Figure 10 illustrates the proportion of deep learning methods. It can be clearly seen that Autoencoder is the most applied deep learning algorithms in documents clustering as at the time of conducting this literature survey; it represented 43% of the studies. However, the focus of these methods to date has been on image data sets [68]. Recurrent represents the second method by 19% in comparison to other techniques followed by Deep Blief and Convolution; they have approximately the same percentages of studies. Finally, Hybrid is a minority research, it is expected to attract more studies in the future.

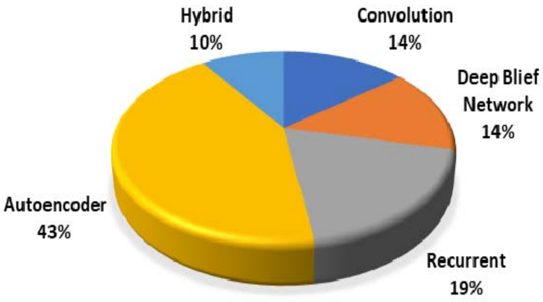

Figure 10: Trend of deep learning methods

Finally, researchers have used a variety of deep learning frameworks and library to achieve the desired results by scanning studies related to clustering documents using deep learning Python ranked as first programming language that the most commonly used followed by Tensorflow and Kersa. Figure 11 shows the number of the studies that used deep learning frameworks and library

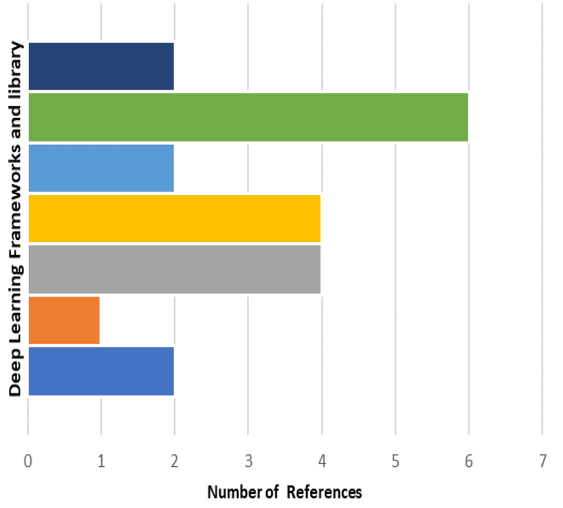

Figure 11: Deep Learning Frameworks and library used in different studies

10 CHALLENGES AND FUTURE RESEARCH DIRECTION

Although deep learning techniques have offered many benefits in document clustering, there are unresolved challenges. This section provides challenges and future research directions with new perspective outlined to facilitate future development of the research area as follow:

✓ *Hybrid and ensemble deep learning* algorithm: In most studies and research, hybrid deep learning algorithms usually achieve distinct results and outperform other individual deep learning algorithms. Hybrid deep learning algorithms take the advantage of each algorithm and overcome the weaknesses; however, according to the previous analysis Figure 10 shows that hybrid algorithms represent a few researches in documents clustering. In future more studies will need to be done in hybrid methods by combining two or more deep learning approaches.

✓ *Extensive research on deep learning*: There is still a research gap where many deep learning algorithms are not utilized or applied in document clustering. For example, Deep Reinforcement Learning has proven successful in many text applications such as [88]and [89] but during our survey, there is no research document clustering using this method or combining it with other deep learning methods; we encourage researchers to focuses on this gap.

✓ *Domain complexity:* As we mentioned in Section 7, there is a wide variation in the document's domain, we need additional research that tests the effectiveness of algorithms by applying them to different types of documents domain, for instance, applying deep learning algorithms in text documents using several data set including business documents, biomedical documents, news documents and so on, and then comparing the results to examine the role, effect of the complexity document domain, and their relationship with clustering documents results.

✓ *Documents languages*: Most of the studies focused on English language text, however, other languages have political and social significance and represent a large percentage of documents that require strong categorization techniques; therefore, we encourage researchers to focus on other local and international languages to cluster these documents and prove the approach effectiveness in future work.

✓ *Create new data set*: Absence of large and free other languages such as powerful and reasonable Arabic datasets is also one of the obstacles [64], researchers can work to create new datasets to solve this problem.

✓ *Experiments in text dataset*: In general, most deep learning experiments on an image data set, while a few attempts have been made on sequential data, e.g., documents [17], we need more experiments for text dataset.

11 CONCLUSIONS

Document clustering is a process of organizing documents by dividing them into several separate groups, each group contains documents with similarly related topics and completely different from the other groups, the importance of clustering is organizing the data, summarizing it through cluster, groupings in the unlabeled data, and identifying the degree of similarity among points. We proposed a survey for deep learning of document clustering to contribute to this growing area of research by providing an important opportunity to understand deep learning methods, dataset, library, programming frameworks and performance metrics used in documents clustering and provide some challenges and suggestion for future research. We encourage researchers to do more research where many deep learning algorithms are not utilized in document clustering.

ACKNOWLEDGEMENT

This research was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University through the Fast-Track Research Funding Program.

REFERENCES

[1] J. Yi, Y. Zhang, X. Zhao, and J. Wan, "A novel text clustering approach using deep-learning vocabulary network," Mathematical Problems in Engineering, vol. 2017, 2017.

[2] S. C. AI. medium. https://medium.com/chat-bots-developers/introduction-to-text-clustering-50d3718ddb01 (accessed.

[3] Q. Liu, J. Wang, D. Zhang, Y. Yang, and N. Wang, "Text Features Extraction based on TF-IDF Associating Semantic," in 2018 IEEE 4th International Conference on Computer and Communications (ICCC), 2018: IEEE, pp. 2338-2343.

[4] K. Sathiyakumari, G. Manimekalai, V. Preamsudha, and M. P. Scholar, "A survey on various approaches in document clustering," International Journal of computer technology and application (IJCTA), vol. 2, no. 5, pp. 1534-1539, 2011.

[5] N. Y. Saiyad, H. B. Prajapati, and V. K. Dabhi, "A survey of document clustering using semantic approach," in 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016: IEEE, pp. 2555-2562.

[6] A. Ghosal, A. Nandy, A. K. Das, S. Goswami, and M. Panday, "A Short Review on Different Clustering Techniques and Their Applications," in Emerging Technology in Modelling and Graphics: Springer, 2020, pp. 69-83.

[7] J. Bae, T. Helldin, M. Riveiro, S. Nowaczyk, M.-R. Bouguelia, and G. Falkman, "Interactive Clustering: A Comprehensive Review," ACM Computing Surveys (CSUR), vol. 53, no. 1, pp. 1-39, 2020.

[8] R. Ibrahim, S. Zeebaree, and K. Jacksi, "Survey on Semantic Similarity Based on Document Clustering," Adv. Sci. Technol. Eng. Syst. J, vol. 4, no. 5, pp. 115-122, 2019.

[9] S. A. Fahad and W. M. Yafooz, "Review on semantic document clustering," International Journal on Contemporary Computer Research (IJCCR), vol. 1, no. 1, pp. 14-30, 2017.

[10] A. Gupta, J. Gautam, and A. Kumar, "A survey on methodologies used for semantic document clustering," in 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017: IEEE, pp. 671-675.

[11] R. Jensi and D. G. W. Jiji, "A survey on optimization approaches to text document clustering," arXiv preprint arXiv:1401.2229, 2014.

[12] K. Mugunthadevi, S. Punitha, M. Punithavalli, and K. Mugunthadevi, "Survey on feature selection in document clustering," International Journal on Computer Science and Engineering, vol. 3, no. 3, pp. 1240-1244, 2011.

[13] C. C. Aggarwal and C. Zhai, "A survey of text clustering algorithms," in Mining text data: Springer, 2012, pp. 77-128.

[14] Y. Xiao, "A Survey of Document Clustering Techniques & Comparison of LDA and moVMF," North Carolina State University, 2010.

[15] H. Patil and R. Thakur, "Document Clustering: A Summarized Survey," in Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications: IGI Global, 2018, pp. 47-64.

[16] E. Aljalbout, V. Golkov, Y. Siddiqui, M. Strobel, and D. Cremers, "Clustering with deep learning: Taxonomy and new methods," arXiv preprint arXiv:1801.07648, 2018.

[17] 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.

[18] A. I. Károly, R. Fullér, and P. Galambos, "Unsupervised clustering for deep learning: A tutorial survey," Acta Polytechnica Hungarica, vol. 15, no. 8, pp. 29-53, 2018.

[19] S. Wang, J. Cai, Q. Lin, and W. Guo, "An Overview of Unsupervised Deep Feature

Representation for Text Categorization," IEEE Transactions on Computational Social Systems, vol. 6, no. 3, pp. 504-517, 2019.

[20] M. Mittal, L. M. Goyal, D. J. Hemanth, and J. K. Sethi, "Clustering approaches for high-dimensional databases: A review," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 3, p. e1300, 2019.

[21] A. Karpathy, "Cs231n convolutional neural networks for visual recognition," Neural networks, vol. 1, 2016.

[22] S.-C. Lo, S.-L. Lou, J.-S. Lin, M. T. Freedman, M. V. Chien, and S. K. Mun, "Artificial convolution neural network techniques and applications for lung nodule detection," IEEE Transactions on Medical Imaging, vol. 14, no. 4, pp. 711-718, 1995.

[23] H. H. Aghdam and E. J. Heravi, "Guide to Convolutional Neural Networks," New York, NY: Springer. doi, vol. 10, pp. 978-3, 2017.

[24] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," Neurocomputing, vol. 234, pp. 11-26, 2017.

[25] Y. Kim, "Convolutional neural networks for sentence classification," arXiv preprint arXiv:1408.5882, 2014.

[26] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," arXiv preprint arXiv:1901.06032, 2019.

[27] S. Saha, "A comprehensive guide to convolutional neural networks—the ELI5 way," Towards Data Science, vol. 15, 2018.

[28] L. Li, X. Li, Y. Yang, and J. Dong, "Indoor tracking trajectory data similarity analysis with a deep convolutional autoencoder," Sustainable cities and society, vol. 45, pp. 588-595, 2019.

[29] Z. Yuhui, G. Gutmann, and K. Akihiko, "Irregular Convolutional Auto-Encoder on Point Clouds," arXiv preprint arXiv:1910.02686, 2019.

[30] A. Ng, "Sparse autoencoder," CS294A Lecture notes, vol. 72, no. 2011, pp. 1-19, 2011.

[31] M. Maggipinto, C. Masiero, A. Beghi, and G. A. Susto, "A Convolutional Autoencoder Approach for Feature Extraction in Virtual Metrology: Paper ID 259," Procedia Manufacturing, vol. 17, pp. 126-133, 2018.

[32] G. Li, D. Han, C. Wang, W. Hu, V. D. Calhoun, and Y.-P. Wang, "Application of deep canonically correlated sparse autoencoder for the classification of schizophrenia," Computer methods and programs in biomedicine, vol. 183, p. 105073, 2020.

[33] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in Proceedings of the 25th international conference on Machine learning, 2008: ACM, pp. 1096-1103.

[34] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," Neural computation, vol. 18, no. 7, pp. 1527-1554, 2006.

[35] G. Fu, "Deep belief network based ensemble approach for cooling load forecasting of air-conditioning system," Energy, vol. 148, pp. 269-282, 2018.

[36] P. Hamel and D. Eck, "Learning features from music audio with deep belief networks," in ISMIR, 2010, vol. 10: Utrecht, The Netherlands, pp. 339-344.

[37] X.-H. Le, H. V. Ho, G. Lee, and S. Jung, "Application of long short-term memory (LSTM) neural network for flood forecasting," Water, vol. 11, no. 7, p. 1387, 2019.

[38] J. Zhou and W. Xu, "End-to-end learning of semantic role labeling using recurrent neural networks," in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 1127-1137.

[39] S. Yoon, H. Yun, Y. Kim, G.-t. Park, and K. Jung, "Efficient transfer learning schemes for personalized language modeling using recurrent neural network," in Workshops at the Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[40] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017: IEEE, pp. 2227-2231.

[41] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[42] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in Fifteenth annual conference of the international speech communication association, 2014.

[43] J. Aungiers, "TIME SERIES PREDICTION USING LSTM DEEP NEURAL NETWORKS," 1st September 2018. [Online]. Available: https://www.altumintelligence.com/articles/a/Time-Series-Prediction-Using-LSTM-Deep-Neural-Networks.

[44] C. Olah, "Understanding lstm networks," 2015. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/#fn1.

[45] F. A. Gers and J. Schmidhuber, "Recurrent nets that time and count," in Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, 2000, vol. 3: IEEE, pp. 189-194.

[46] K. Yao, T. Cohn, K. Vylomova, K. Duh, and C. Dyer, "Depth-gated LSTM," arXiv preprint arXiv:1508.03790, 2015.

[47] J. Koutnik, K. Greff, F. Gomez, and J. Schmidhuber, "A clockwork rnn," arXiv preprint arXiv:1402.3511, 2014.

[48] M. Kampffmeyer, S. Løkse, F. M. Bianchi, L. Livi, A.-B. Salberg, and R. Jenssen, "Deep divergence-based approach to clustering," Neural Networks, vol. 113, pp. 91-101, 2019.

[49] J. Xu et al., "Short text clustering via convolutional neural networks," in Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, 2015, pp. 62-69.

[50] J. Xu, B. Xu, P. Wang, S. Zheng, G. Tian, and J. Zhao, "Self-taught convolutional neural networks for short text clustering," Neural Networks, vol. 88, pp. 22-31, 2017.

[51] Z. Wang, H. Mi, and A. Ittycheriah, "Semi-supervised clustering for short text via deep representation learning," arXiv preprint arXiv:1602.06797, 2016.

[52] G. Chen, "Deep learning with nonparametric clustering," arXiv preprint arXiv:1501.03084, 2015.

[53] V. Jindal, "A personalized Markov clustering and deep learning approach for Arabic text categorization," in Proceedings of the ACL 2016 Student Research Workshop, 2016, pp. 145-151.

[54] A. A. Frolov, D. Husek, and P. Y. Polyakov, "Recurrent-neural-network-based Boolean factor analysis and its application to word clustering," IEEE transactions on neural networks, vol. 20, no. 7, pp. 1073-1086, 2009.

[55] S. Shah and X. Luo, "Comparison of deep learning based concept representations for biomedical document clustering," in 2018 IEEE EMBS international conference on biomedical & health informatics (BHI), 2018: IEEE, pp. 349-352.

[56] B. Wang, W. Liu, Z. Lin, X. Hu, J. Wei, and C. Liu, "Text clustering algorithm based on deep representation learning," The Journal of Engineering, vol. 2018, no. 16, pp. 1407-1414, 2018.

[57] T. Duan, Q. Lou, S. N. Srihari, and X. Xie, "Sequential embedding induced text clustering, a non-parametric bayesian approach," in Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2019: Springer, pp. 68-80.

[58] J. Park, C. Park, J. Kim, M. Cho, and S. Park, "ADC: Advanced document clustering using contextualized representations," Expert Systems with Applications, vol. 137, pp. 157-166, 2019.

[59] X. Guo, X. Liu, E. Zhu, and J. Yin, "Deep clustering with convolutional autoencoders," in International Conference on Neural Information Processing, 2017: Springer, pp. 373-382.

[60] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in International conference on machine learning, 2016, pp. 478-487.

[61] X. Guo, L. Gao, X. Liu, and J. Yin, "Improved deep embedded clustering with local structure preservation," in IJCAI, 2017, pp. 1753-1759.

[62] B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong, "Towards k-means-friendly spaces: Simultaneous deep learning and clustering," in Proceedings of the 34th International Conference on Machine Learning-Volume 70, 2017: JMLR. org, pp. 3861-3870.

[63] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, "Variational deep embedding: An unsupervised and generative approach to

clustering," arXiv preprint arXiv:1611.05148, 2016.

[64] K. Tian, S. Zhou, and J. Guan, "Deepcluster: A general clustering framework based on deep learning," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2017: Springer, pp. 809-825.

[65] U. Shaham, K. Stanton, H. Li, B. Nadler, R. Basri, and Y. Kluger, "Spectralnet: Spectral clustering using deep neural networks," arXiv preprint arXiv:1801.01587, 2018.

[66] D. Xu et al., "Deep co-clustering," in Proceedings of the 2019 SIAM International Conference on Data Mining, 2019: SIAM, pp. 414-422.

[67] J. Kim, J. Yoon, E. Park, and S. Choi, "Patent document clustering with deep embeddings," Scientometrics, pp. 1-15, 2020.

[68] M. Gong, C. Yao, Y. Xie, and M. Xu, "Semi-supervised Network Embedding with Text Information," Pattern Recognition, p. 107347, 2020.

[69] A. Hadifar, L. Sterckx, T. Demeester, and C. Develder, "A self-training approach for short text clustering," in Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), 2019, pp. 194-199.

[70] X.-H. Dang, R. Akella, S. Bahrami, V. Sheinin, and P. Zerfos, "Unsupervised Threshold Autoencoder to Analyze and Understand Sentence Elements," in 2018 IEEE International Conference on Big Data (Big Data), 2018: IEEE, pp. 3267-3276.

[71] K. Lang, "Newsweeder: Learning to filter netnews," in Machine Learning Proceedings 1995: Elsevier, 1995, pp. 331-339.

[72] C. Baker and S. P. Jones, Encyclopedia of bilingualism and bilingual education. Multilingual Matters, 1998.

[73] U.Nations. https://www.un.org/en/sections/about-un/official-languages/ (accessed.

[74] https://www.definitions.net/definition/REGIONAL+LANGUAGE (accessed.

[75] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[76] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in International conference on machine learning, 2014, pp. 1188-1196.

[77] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532-1543.

[78] M. E. Peters et al., "Deep contextualized word representations," arXiv preprint arXiv:1802.05365, 2018.

[79] Y. Jia et al., "Caffe: Convolutional architecture for fast feature embedding," in Proceedings of the 22nd ACM international conference on Multimedia, 2014: ACM, pp. 675-678.

[80] [Online]. Available: https://keras.io/.

[81] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit," Information Processing & Management, vol. 57, no. 2, p. 102034, 2020.

[82] M. Afreen and S. Badugu, "Document Clustering Using Different Unsupervised Learning Approaches: A Survey," in Advances in Decision Sciences, Image Processing, Security and Computer Vision: Springer, 2020, pp. 619-629.

[83] C. Luo, Y. Li, and S. M. Chung, "Text document clustering based on neighbors," Data & Knowledge Engineering, vol. 68, no. 11, pp. 1271-1288, 2009.

[84] H. Wan, B. Ning, X. Tao, and J. Long, "Research on Chinese Short Text Clustering Ensemble via Convolutional Neural Networks," in Artificial Intelligence in China: Springer, 2020, pp. 622-628.

[85] H. Bunyamin, S. Novianti, and L. Sulistiani, "Topic Clustering and Classification on Final Project Reports: a Comparison of Traditional and Modern Approaches," IAENG International Journal of Computer Science, vol. 46, no. 3, 2019.

[86] Z. H. Kilimci and S. Akyokuş, "The Analysis of Text Categorization Represented With Word Embeddings Using Homogeneous Classifiers," in 2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA), 2019: IEEE, pp. 1-6.

[87] I. S. Dhillon, S. Mallela, and R. Kumar, "Enhanced word clustering for hierarchical text classification," in Proceedings of the eighth ACM SIGKDD international

conference on Knowledge discovery and data mining, 2002, pp. 191-200.

[88] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in Natural language processing and text mining: Springer, 2007, pp. 9-28.

[89] M. Yang, Q. Qu, Y. Shen, K. Lei, and J. Zhu, "Cross-domain aspect/sentiment-aware abstractive review summarization by combining topic modeling and deep reinforcement learning," Neural Computing and Applications, pp. 1-13, 2018.