International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.7930
Article Received: 02 May 2023 Revised: 22 June 2023 Accepted: 28 July 2023

Harnessing Deep Learning Techniques for Text Clustering and Document Categorization

Rama Krishna Paladugu[1], Gangadhara Rao Kancherla[2]
[1] Research Scholar, Department of Computer Science and Engineering,
Acharya Nagarjuna University, Guntur 522510, India.
Email:mails4prk@gmail.com
[2] Professor, Department of Computer Science and Engineering,
Acharya Nagarjuna University, Guntur 522510, India.
Email:kancherla123@gmail.com

Abstract—This research paper delves into the realm of deep text clustering algorithms with the aim of enhancing the accuracy of document classification. In recent years, the fusion of deep learning techniques and text clustering has shown promise in extracting meaningful patterns and representations from textual data. This paper provides an in-depth exploration of various deep text clustering methodologies, assessing their efficacy in improving document classification accuracy. Delving into the core of deep text clustering, the paper investigates various feature representation techniques, ranging from conventional word embeddings to contextual embeddings furnished by BERT and GPT models.By critically reviewing and comparing these algorithms, we shed light on their strengths, limitations, and potential applications. Through this comprehensive study, we offer insights into the evolving landscape of document analysis and classification, driven by the power of deep text clustering algorithms.Through an original synthesis of existing literature, this research serves as a beacon for researchers and practitioners in harnessing the prowess of deep learning to enhance the accuracy of document classification endeavors.

Keywords- Contextual embeddings, Semantic embeddings, Document classification, Natural language processing, Deep text clustering and Deep learning.

I. INTRODUCTION

In the realm of natural language processing (NLP) [1], the semantic fusion of deep learning techniques and text clustering algorithms [2] has garnered significant attention, promising to reshape the landscape of document analysis and classification [3 and 4]. This introductory section lays the groundwork for our research paper, elucidating the contextual backdrop, the driving motivation, and the carefully delineated trajectory of our exploration. By unraveling the threads of research, we attempt to write texture this study that encompasses the dimensions, challenges, and innovations inherent in the synergetic fusion of deep learning techniques and text clustering algorithms.

1.1 Background and Motivation: The exponential growth of textual data across diverse domains has underscored the imperative to decipher its latent semantics for efficient information retrieval, categorization, and decision-making [5]. Traditional text clustering methodologies [6 to 10] have demonstrated utility in organizing documents into coherent groups, but their efficacy in capturing intricate semantic relationships remains constrained. This inadequacy prompted the emergence of deep learning paradigms [11 and 12] that excel in extracting intricate patterns and hierarchies from unstructured text, catapulting the field into a new era of text analysis.

1.2 Problem Statement: Amid this backdrop, a key challenge [13] emerges is “How can deep text clustering algorithms be harnessed to augment the accuracy of document classification?”. This essential investigation underscores our research's pivotal aim—to systematically analyze the probability of these algorithms in elevating the precision of document categorization. By acknowledging the limitations of traditional clustering approaches [4, 6 and 13] and the flourishing capabilities of deep learning models, we seek to bridge the gap between conventional methodologies and the demands of modern document analysis.

1.3 Research Objectives: Aligned with the aforementioned problem statement, this study concentrates on a multifaceted journey with distinct objectives. Firstly, we aim to comprehensively study and synthesize the existing body of literature concerning deep text clustering algorithms and their application in document classification scenarios. Secondly, our research endeavors to critically assess the performance of these algorithms in comparison to traditional clustering techniques, highlighting the nuances and advantages that stem from their deep learning substructures.

1.4 Scope and Organization: It is essential to define the scope of this research to clarify its boundaries. Our research focuses primarily on the examination of various deep text clustering algorithms and their implications for document classification accuracy [13]. While we acknowledge the broader domain of NLP [1], our primary emphasis rests on the synergy between deep learning and text clustering [14]. We concentrate on algorithms that integrate neural architectures with text clustering [15], exploring the crossroads where these methods intersect. This approach is not restrictive but guides us towards a comprehensive analysis. By honing in on this intersection, we reveal how deep learning techniques can enhance document classification precision [9].

Our pointed focus allows us to delve into the essentials of deep text clustering algorithms in the context of document classification [11, 13, 14 and 15]. It enables us to uncover patterns that might be obscured with a broader scope. However, this focus doesn't stand alone; it's a lens that magnifies the point where deep learning and text clustering converge.As deep learning continues to infiltrate NLP, our research establishes a dedicated path within this complex landscape. We delve into the connections and outcomes that arise from merging these domains. Our exploration sets out on a journey that aligns with NLP's broader objectives [1] while remaining rooted in the transformative potential of deep text clustering algorithms.

The paper's subsequent sections are appropriately structured to fulfill our research objectives. We delve into the landscape of text clustering algorithms, explore the role of deep learning in this context, detail our methodology, present a comparative analysis, discuss our findings, and culminate with a comprehensive conclusion.

II. LITERATURE REVIEW

2.1 Document Classification and Text Clustering

In the dynamic field of natural language processing (NLP) [1], the domains of document classification and text clustering stand as pivotal pillars, collectively driving advancements in understanding and organizing textual data. Document classification involves categorizing documents into predefined classes or categories based on their content [16]. This process enables effective information retrieval, content organization, and decision-making. On the other hand, text clustering involves grouping similar documents together based on shared characteristics, without predefined categories. Both document classification and text clustering plays indispensable roles in managing and comprehending large volumes of textual data [17].

Document classification operates on the fundamental concept of labeling documents with relevant categories [18 and 19]. This process entails the utilization of various features, such as keywords, semantic information, and structural attributes, to categorize documents accurately [20]. The categorization process often involves supervised learning techniques, where machine learning models learn from labeled training data to predict the classes of new, unseen documents [8]. In contrast, text clustering aims to discover hidden patterns and inherent structures within a corpus of text [15]. Unlike document classification, text clustering is an unsupervised learning task, where the objective is to group similar documents together, often revealing latent semantic relationships that may not be apparent on the surface.

The accuracy of document classification [16] carries significant implications for information organization, search efficiency, and decision support systems. In domains such as information retrieval, accurate classification ensures that relevant documents are retrieved promptly in response to user queries, enhancing user satisfaction. Additionally, accurate document classification aids in automating tasks that require content categorization, such as news categorization, sentiment analysis, and spam detection [21]. Furthermore, in industries where compliance and regulatory adherence are critical, precise classification facilitates the proper management of sensitive documents.

While traditional document classification methods [6, 7, 8, 11, 14 and 17] have provided valuable insights, the advent of deep learning has introduced new dimensions to these tasks. Deep learning models, with their capacity to extract intricate patterns and representations from text, have shown promise in enhancing classification accuracy [2, 13, 15 and 20]. However, the integration of deep learning with text clustering methodologies is a relatively unexplored frontier, one that holds the potential to reshape how document classification and text clustering are synergized for improved accuracy and meaningful insights [20]. This research paper explores into this exciting study, where deep contextual embedding models and clustering algorithms intersect to propel the accuracy of document classification to new heights.

2.2 Traditional Clustering Methods

In the realm of text analysis and document classification, traditional clustering methods stand as cornerstones for organizing and understanding textual data. As we embark on this exploration of deep text clustering algorithms, it's imperative to comprehend the environment provided by these established techniques. These traditional methods lay a foundation that contextualizes the innovative deep learning approaches and their implications for document classification.

2.2.1 K-means Clustering:K-means clustering, a widely adopted technique in data analysis, finds its relevance in the context of deep text clustering and document classification. At its core, K-means aims to partition a dataset into distinct clusters, where each data point belongs to the cluster with the nearest mean value. In the landscape of text, this translates to grouping similar documents together based on shared features or characteristics. However, while K-means effectively identifies clusters, its rigid partitioning may struggle to capture the intricate semantic relationships inherent in textual content. The emerging fusion of deep learning with clustering techniques holds the promise of alleviating this limitation by integrating semantic understanding into cluster formation.

In their innovative study, Rashid et al. (2019) [22] introduce a novel topic modeling technique tailored for text mining within biomedical text corpora. Their proposed model combines hybrid inverse document frequency (IDF) with fuzzy K-means clustering. This approach overcomes the challenge of effectively organizing complex biomedical text by leveraging IDF to extract salient features and employing fuzzy K-means for nuanced cluster formation. The novelty stems from the fusion of hybrid IDF and fuzzy K-means, providing a robust solution for enhancing the organization and understanding of biomedical data. In their pioneering work, Abasi et al. (2020) [23] present an innovative hybrid approach that merges the multi-verse optimizer (MVO) with K-means for text document clustering. This model effectively addresses the challenge of organizing large volumes of textual data by leveraging the diversity exploration of MVO and the cluster formation capabilities of K-means. The novelty lies in the integration of MVO and K-means, providing a synergistic solution that contributes to the advancement of deep text clustering methods for improved document classification outcomes.

Amer and Abdalla (2020) [24] introduce a new set theory-based similarity measure designed to address the challenge of text clustering and classification. Their innovative approach employs set operations to quantify textual similarity, effectively overcoming the intricacies of capturing semantic relationships within documents. By offering a robust solution to the complexities of text analysis, this proposed model contributes to enhancing document clustering and classification accuracy. The novelty lies in the utilization of set theory principles, offering a fresh perspective on deep text clustering that empowers more accurate and meaningful document categorization. Jo (2008) [25] presents an inventive approach by introducing an inverted index-based modification to the K-means algorithm, tailored for text clustering challenges. The proposed model effectively addresses the intricacies of text clustering by leveraging inverted indexes to enhance the clustering process's efficiency. This modification overcomes the limitations of traditional K-means in handling large textual datasets. The novelty resides in the integration of an inverted index mechanism, yielding accelerated clustering performance and fostering improved accuracy in document classification.

Limitations in K-means Text Clustering: Despite its popularity, K-means clustering has certain limitations, particularly when applied to text data in the context of deep text clustering and document classification. While K-means is adept at partitioning datasets into clusters based on mean values, its rigidity in partitioning can hinder its ability to capture the nuanced semantic relationships inherent in textual content. In the landscape of text analysis, where documents often possess intricate patterns and subtle contextual meanings, K-means' straightforward approach may result in clusters that fail to adequately reflect the subtleties within the data.

2.2.2 Hierarchical Clustering: Hierarchical clustering, another traditional approach, offers a different perspective within the context of deep text clustering and document classification. Unlike K-means, hierarchical clustering creates a tree-like structure of nested clusters, providing a hierarchical representation of data relationships. In the context of textual data, this hierarchical approach can capture both global and local patterns, potentially enhancing the discovery of meaningful thematic groups within documents. However, traditional hierarchical clustering methods might struggle with the nuanced semantics of natural language, leaving room for the infusion of deep learning to augment its accuracy and adaptability to textual content.

Pappagari et al. (2019) [26] present an innovative model, Hierarchical Transformers, catering to the challenge of classifying long documents. By integrating hierarchical structures within transformer architectures, this approach addresses the limitations in accurately classifying extensive textual content.The novelty lies in the fusion of transformer capabilities with hierarchical structures, offering an advanced solution for document classification in the context of deep text clustering.Stein et al. (2019) [27] introduce an insightful investigation into hierarchical text classification, leveraging the power of word embeddings. Their proposed model employs hierarchical structures to navigate the complexities of text categorization, effectively addressing challenges in organizing large textual datasets.By incorporating word embeddings, the approach captures semantic nuances, bolstering classification accuracy.Meng et al. (2019) [28] present a groundbreaking approach in weakly-supervised hierarchical text classification, innovatively addressing the challenge of limited labeled data. Their model navigates the complexities of document classification by leveraging hierarchical structures and weak supervision. This approach overcomes the limitations posed by scarce labeled data, enhancing classification accuracy.

Limitations in Hierarchical Clustering: While hierarchical clustering provides a hierarchical representation of data relationships, offering a valuable perspective for deep text clustering and document classification, it has certain limitations that deserve consideration. Despite its potential to capture both global and local patterns within textual data, traditional hierarchical clustering methods may grapple with the nuanced semantics inherent in natural language. The intricacies of language, such as polysemy and context-dependent meanings, can pose challenges for hierarchical clustering's ability to accurately capture semantic relationships within documents.

2.2.3 DBSCAN: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) introduces yet another dimension to the spectrum of traditional clustering methods in the context of our exploration. DBSCAN identifies clusters based on density connectivity, effectively capturing areas of high-density points in the data space. In the context of text clustering and document classification, DBSCAN's capacity to uncover clusters of varying shapes and sizes has potential benefits. However, this method's sensitivity to parameter settings and susceptibility to noise might limit its performance when applied directly to textual data.

Cretulescu et al. (2019) [29] introduce an innovative application of the DBSCAN algorithm tailored for document clustering, addressing the intricacies of organizing textual data. Their model adapts the DBSCAN algorithm, originally designed for spatial data, to tackle the challenges posed by document organization. This approach overcomes limitations in traditional document clustering methods by leveraging density-based clustering to identify clusters of varying shapes and sizes. Liu and Yang (2022) [30] contribute to the domain of deep text clustering and document classification with their research on web text clustering. Their proposed model leverages an optimized version of the DBSCAN algorithm to address the challenges of effectively clustering web-based textual data. By tailoring DBSCAN to the specific requirements of web text clustering, the model overcomes the limitations posed by noise and varying data densities.

Mohammed et al. (2020) [31] contribute to the advancement of deep text clustering and document classification with their research on semantic document clustering. Their proposed model integrates GloVe word embeddings with the DBSCAN algorithm to address the challenge of effectively clustering textual data based on semantic meanings. By incorporating word embeddings, the model overcomes limitations in traditional methods by capturing intricate semantic relationships between words.

Limitations of DBSCAN: While DBSCAN presents a promising approach in the context of text clustering and document classification, it is not exempt from certain limitations that deserve consideration. Despite its capacity to uncover clusters of varying shapes and sizes based on density connectivity, this method's performance can be sensitive to parameter settings. Inaccurate parameter choices may lead to suboptimal cluster formations, hindering the precision and reliability of cluster assignments, especially when applied directly to textual data with its inherent complexity.Furthermore, DBSCAN's susceptibility to noise can pose challenges in the context of deep text clustering.

While traditional clustering methods provide valuable foundational insights, their limitations become apparent when handling the intricate nuances of textual content. The integration of deep learning techniques presents a transformative opportunity to address these limitations and elevate the accuracy and effectiveness of document classification through advanced contextual embeddings and innovative cluster formation methods. As we navigate through the landscape of deep text clustering, these traditional methods serve as essential benchmarks, guiding us toward a deeper understanding of how modern techniques can enhance the classification of textual data.

2.3 Deep Learning in Text Clustering

The integration of deep learning techniques with text clustering has ushered in a new era of exploration in document analysis and classification. This subsection delves into key approaches that leverage the power of deep learning to enhance text clustering outcomes, spanning Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Autoencoders, and Self-Organizing Maps (SOMs).

2.3.1 CNNs for Text Clustering: CNNs renowned for their effectiveness in image analysis, have found application in text clustering as well. By treating textual data as a form of image, where words are represented in a matrix format, CNNs extract meaningful features using convolutional layers. This technique can capture local patterns and relationships within textual data, making it suitable for tasks like sentence or document embeddings. However, while CNNs excel at extracting features from fixed-size windows, they might struggle with capturing long-range dependencies that are prevalent in text.

Widiastuti (2019) [32] presents an impactful contribution to the realm of deep text clustering and document classification through the application of CNNs. Their proposed model capitalizes on CNNs' prowess in image analysis to process textual data, effectively addressing challenges in text mining and natural language processing. By leveraging CNNs' ability to capture local patterns and relationships, the model overcomes limitations in traditional methods, enhancing the accuracy of text mining tasks. Aich et al. (2019) [33] contribute significantly to the domain of deep text clustering and document classification by introducing a CNN based model tailored for web-based text classification. Their proposed model harnesses the capabilities of CNNs to effectively process and classify textual data sourced from the web. This innovative approach overcomes challenges associated with the vast and diverse nature of web-based text. By capitalizing on CNNs' capacity to capture local patterns and relationships, the model enhances classification accuracy.

Wang et al. (2020) [34] make a significant contribution to the field of deep text clustering and document classification by proposing a Graph Attention Convolutional Neural Network (GCN) model for predicting chemical poisoning in honey bees. Their innovative approach leverages GCNs to analyze and classify textual data related to honey bee health and chemical exposure. This model effectively addresses the challenge of predicting complex biological outcomes using textual information.

Limitations of CNNs in Text Clustering:While Convolutional Neural Networks (CNNs) have demonstrated effectiveness in image analysis and are increasingly applied to text clustering, they come with certain limitations when dealing with textual data. CNNs treat text as an image, which allows them to extract meaningful features using convolutional layers. This approach is particularly adept at capturing local patterns and relationships, making it suitable for tasks like sentence or document embeddings. However, the inherent nature of text differs from images in that it often contains long-range dependencies and intricate semantic relationships.

2.3.2 RNNs in Clustering: RNNs offer a dynamic approach to text clustering, allowing the modeling of sequential data. RNNs, with their ability to maintain memory of past inputs, are well-suited for capturing contextual dependencies in text. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) variations of RNNs mitigate the vanishing gradient problem, enabling them to capture longer-range dependencies. However, RNNs have limitations in handling excessively long sequences and might struggle with preserving information over extended distances.

Skrlj et al. (2019) [35] make a notable contribution to the realm of deep text clustering and document classification through their proposed model, a Semantics-Aware Recurrent Neural Architecture. This model tackles the challenge of robust text classification by leveraging recurrent neural networks (RNNs) enriched with semantic understanding. By incorporating semantic information, the model overcomes limitations associated with the complex and varying nature of textual data. Murthy et al. (2020) [36] contribute to the field of deep text clustering and document classification by introducing a model that employs Long Short-Term Memory (LSTM) networks for text-based sentiment analysis. Their proposed approach utilizes LSTM's ability to capture sequential dependencies within textual data, enhancing sentiment analysis accuracy. This model effectively addresses challenges in understanding and categorizing sentiment-laden textual content.

Limitations of RNNs in Text Clustering:While RNNs offer a promising approach to capturing contextual dependencies in text, they come with inherent limitations that should be considered. RNNs excel in modeling sequential data by maintaining memory of past inputs, which suits the intricacies of textual content. Variations like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) further alleviate the vanishing gradient problem, enabling them to capture longer-range dependencies effectively. However, despite these advantages, RNNs face challenges in handling excessively long sequences. As the distance between relevant elements within a sequence increases, RNNs may struggle to preserve information over extended distances, resulting in the potential loss of vital context. This limitation could adversely impact the accurate understanding and representation of complex textual relationships.

2.3.3 Auto encoders for Text Clustering: Autoencoders, a form of unsupervised deep learning, aim to reconstruct input data by encoding it into a latent space and then decoding it back into the original space. In text clustering, autoencoders can learn compressed representations of textual data. This technique is particularly beneficial for dimensionality reduction and noise reduction, enhancing the efficiency of subsequent clustering algorithms. However, the performance of autoencoders heavily relies on the choice of architecture and the quality of the encoded representations.

Hosseini and Varzaneh (2022) [37] contribute to the advancement of deep text clustering and document classification with their proposed model that employs Stacked AutoEncoders. Their approach addresses the challenge of intricate text clustering by utilizing AutoEncoders to learn effective feature representations from textual data. By capturing latent patterns and representations, this model overcomes complexities in organizing and categorizing textual information.Yin et al. (2021) [38] present a notable contribution to the domain of deep text clustering and document classification by proposing a representation learning approach tailored for short text clustering. Their model addresses the challenge of effectively clustering short textual content by employing advanced representation learning techniques. This approach overcomes the limitations of traditional methods in capturing the nuances and semantic relationships within concise text segments.

Limitations of Autoencoders in Text Clustering: Autoencoders, as a powerful unsupervised deep learning technique, offer valuable advantages in text clustering by learning compact representations of textual data through encoding and decoding. This capability aids in dimensionality reduction and noise reduction, thus improving the efficiency of downstream clustering methods. However, the efficacy of autoencoders is closely tied to architectural decisions and the quality of the encoded representations. Despite their potential, autoencoders have certain limitations. The selection of appropriate architecture demands expertise, and suboptimal choices can hinder their performance. Additionally, the quality of encoded features is crucial for subsequent clustering, as inaccurate or irrelevant features could lead to poor cluster quality and classification accuracy.

III. DEEP TEXT CLUSTERING ALGORITHMS

Deep Text Clustering Algorithms [15, 18 and 20] represent a fusion of advanced deep learning techniques with in the field of text clustering and document classification. Unlike traditional methods, deep text clustering algorithms delve into the semantic nuances of text, capturing contextual relationships that are often challenging for conventional algorithms. By leveraging these deep learning models, the algorithms aim to enhance the accuracy and efficiency of clustering tasks, enabling better organization and categorization of diverse textual content. Through the utilization of neural architectures designed to comprehend patterns within text, deep text clustering algorithms offer a novel approach to document analysis, empowering researchers and practitioners to uncover hidden insights and optimize classification outcomes.

3.1 Feature Representation and Embeddings

In the realm of Deep Text Clustering Algorithms, effective representation of textual data plays a pivotal role in achieving accurate document classification. This subsection delves into advanced techniques for feature extraction [15] and embedding that empower these algorithms to capture intricate semantic relationships within text.

3.1.1 Word Embeddings (Word2Vec, GloVe): Word embeddings are foundational to deep text clustering, translating words into continuous vector representations that encode semantic meaning. Approaches like Word2Vec [39 and 40] and GloVe [41 and 42] are widely employed to generate these embeddings. Word2Vec learns word representations by predicting words based on their context, while GloVe constructs embeddings by considering word co-occurrence statistics. These embeddings enable algorithms to comprehend semantic relationships, aiding in tasks such as identifying synonymous terms and contextual similarities within documents.

Word2Vec embeddings: For a given word, the Word2Vec model [39] aims to maximize the likelihood of predicting context words within a specific window:

$$Maximize\left(\sum \log P\left(\frac{V_c}{V_w}\right)\right) \qquad (1)$$

Where word is V_w , context is V_C and P(Context | Word) is a softmax function over the dot products of the word and context vectors.

GloVe embeddings: GloVe constructs word embeddings [41] by factoring the co-occurrence matrix of words to capture the relationship between words. Let's denote the scalar product of the vectors for word i and context j as X{ij}, and the frequency of co-occurrence of word i and context j as F{ij}.

$$Minimize \sum\left(X_{\{i,j\}} - \log(F_{\{i,j\}})\right)^2 \qquad (2)$$

This optimization objective aims to find word vectors that best represent the co-occurrence information between words and contexts.

3.1.2 Contextual Embeddings (BERT, GPT):

Contextual embeddings [43 and 44], represented by models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), take feature representation a step further by considering the entire context of a word within a sentence. These models use attention mechanisms to capture both left and right context, yielding highly contextualized embeddings. Contextual embeddings are valuable for capturing nuances, such as polysemy and syntactic structure, within textual content. This deeper understanding enhances the ability of deep text clustering algorithms to distinguish between documents with similar words but distinct meanings.

BERT contextual embeddings: BERT [45] generates contextual embeddings by considering the bidirectional context of each word in a sentence. For a given sentence with N words, let E1, E2, ..., EN represent the contextual embeddings of each word.The contextual embeddings are obtained through multi-layer self-attention mechanisms that capture the relationships between words in both directions.

$$E_i = BERT\_CTEM\ (W_i) \qquad (3)$$

These embeddings, E_i, capture the nuanced meaning of the word (W_i) within its surrounding context BERT_CTEM, enhancing its representation for clustering tasks. The utilization of such embeddings addresses the challenge of representing words more comprehensively in the context of their usage, resulting in improved accuracy for downstream clustering processes.

DOI: https://doi.org/10.17762/ijritcc.v11i8.7930
***Article Received**: 02 May 2023 **Revised**: 22 June 2023 **Accepted**: 28 July 2023*

GPT contextual embeddings: GPT[46] generates contextual embeddings by employing transformer architecture to capture the contextual relationships between words in a sentence. Let G1, G2, ..., GN represent the contextual embeddings of words in a given text sequence with N words, generated using the GPT architecture. These embeddings are derived from the transformer's multi-layer attention mechanisms that capture the contextual relationships between words in a bidirectional manner.

$$G_i = GPT\_CTEM\ (W_i) \qquad (4)$$

These contextual embeddings GPT_CTEM, encapsulate the intricate semantics for G_i and contextual significance of each word W_i within the text. These embeddings can serve as enhanced feature representations for subsequent text clustering and document classification tasks, capturing the intricate nuances of language usage. In the context of text clustering, the GPT-based contextual embeddings G1 to GNfor the words in a text sequence can be aggregated to form an aggregated representative embedding 'T' for the entire text as follows.

$$T = AGR\_EMBED(G_1, G_2, G_3 ..., G_N) \qquad \textbf{(5)}$$

In this aggregated embedding (AGR_EMBED), the T captures the holistic semantic representation of the text, which can then be utilized as input for clustering algorithms.

By integrating these feature representation and embedding techniques, Deep Text Clustering Algorithms can effectively capture the intricacies of language semantics, facilitating improved document classification and clustering accuracy. These techniques offer a holistic approach to deep text analysis, allowing algorithms to navigate the complexities of natural language and unlock deeper insights from textual data.

3.1.3 Self-attention mechanisms:

Self-attention mechanisms [47 and 48] have emerged as a pivotal ingredient in the realm of deep text clustering [37], particularly within the context of semantic understanding [31 and 35] and document classification. These mechanisms introduce a dynamic way to capture intricate relationships within text, revolutionizing the way we approach feature extraction and clustering in the realm of natural language processing.In the paradigm of deep semantic text clustering[35], self-attention operates by assigning distinct roles to each word or token in a given document. This entails treating each element as a query, a key, and a value. Let's consider a sequence of words in a document represented as W = [w1, w2, ..., wn].Each word wiis projected into three spaces: Query (Q), Key (K), and Value (V), forming matrices as Q=[q1,q2,…qn], K=[k1,k2,…kn] and V=[v1, v2,…,vn]. The attention score between each query and key is calculated using a similarity function (e.g., dot product or scaled dot product):

$$Attention(Q,K) = Soft\max\left(\frac{QK^T}{\sqrt{d}}\right) \qquad \textbf{(6)}$$

The attention scores are used to weight the values in the Value matrix to obtain the contextually enriched representation:

$$CTX\ \_REP = Attention\ (Q,K)V \qquad \textbf{(7)}$$

In the context of semantic understanding and document classification, these equations collectively capture the dynamic process of assigning weights to words based on their contextual relevance. This enables the model to extract and emphasize crucial information, essential for accurate semantic interpretation and improved document categorization. By assessing the interactions between these components, self-attention calculates attention scores that reflect the importance of different words in relation to one another. The outcome is a recalibrated representation for each word, enriched with the collective contextual information present within the entire document.

When integrated into deep text clustering algorithms [37], such as transformers [16], self-attention mechanisms [47 and 48] profoundly impact the quality of the extracted features. By weighting words based on their contextual significance, these mechanisms empower models to discern complex patterns, dependencies, and underlying semantics within the text. This is particularly crucial for tackling the intricacies of semantic clustering and document classification, where capturing nuanced relationships and subtleties in meaning is paramount.

IV. METHODOLOGY

4.1 Implementation of Deep Text Clustering

This customized abstract algorithm outlines the intricate process of deep text clustering [37] and document classification, encompassing data preparation, semantic embedding, clustering, evaluation, and interpretation. It allows for flexibility in selecting techniques based on the specific goals and characteristics of the dataset.

Algorithm for Deep Text Clustering

Step 1: Data Preprocessing-Clean and prepare the textual dataset (D) for analysis. Tokenize the documents and remove stop words to enhance data quality.

Step 2: Embedding Generation-Utilize the chosen embedding technique (E) to create semantic embeddings for each document. These embeddings capture the contextual meaning of words and phrases.

Step 3: Dimensionality Reduction (Optional)-If desired, apply dimensionality reduction techniques to simplify data representation.

Step 4: Clustering-Implement the selected clustering method (C) on the embeddings or the reduced dimensions. The algorithm groups similar documents based on their semantic content.

Step 5: Cluster Analysis-Evaluate cluster quality using appropriate metrics (e.g., cluster purity, NMI, ARI). Interpret clusters to uncover relationships within the dataset.

Step 6: Visualization (Optional)-visualize the clustered documents in a lower-dimensional space.

Step 7: Results Interpretation- Analyze and interpret the clustering results to identify thematic patterns.

The block diagram of the abstract deep text clustering algorithm is presented in Figure-1 with a series of steps involved in processing. In the realm of deep text clustering and document classification, the initial phase encompasses the careful collection of pertinent textual data. This phase is helpful in shaping the subsequent stages of analysis. The curated data is subjected to preprocessing techniques [49] tailored to enhance its quality and prepare it for comprehensive examination. Vital steps such as text cleaning, tokenization to break down textual content into individual units, the removal of stopwords to eliminate common, non-informative words, and the application of stemming or lemmatization to reduce words to their root forms, are executed. These measures collectively contribute to the creation of a structured and harmonized dataset. The overarching objective here is to foster a foundation of data that is both refined and poised for meaningful exploration within the context of deep text clustering and document classification.

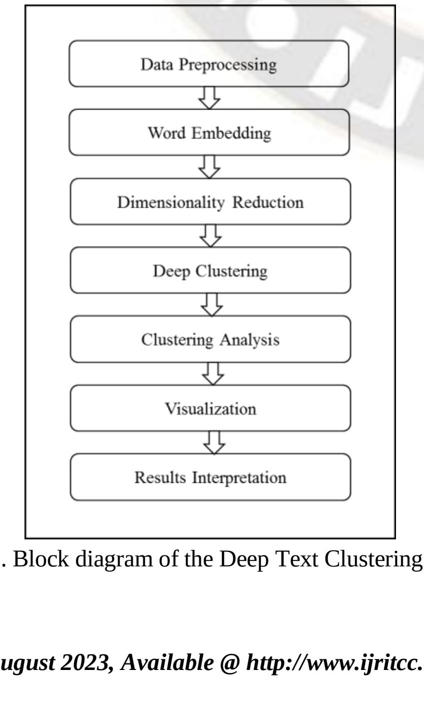

Figure 1. Block diagram of the Deep Text Clustering Algorithms

In the context of advancing deep text clustering and document classification [16], the implementation phase constitutes a pivotal juncture. Here, the spotlight is directed towards the judicious application of cutting-edge deep text clustering algorithms onto the preprocessed dataset. This process involves the integration of techniques like Word2Vec, GloVe [31], BERT [45], GPT [46], and self-attention mechanisms [48], which collectively serve as catalysts in generating intricate semantic embeddings. These embeddings, constructed with a keen focus on capturing the inherent meaning embedded within the textual content, serve as profound representations. Their role is to encapsulate the nuanced contextual essence of the text, thereby laying a solid groundwork for subsequent stages of clustering. By imbuing the dataset with these profound embeddings, the implementation phase sets the stage for sophisticated clustering methodologies to extract meaningful patterns and relationships from the enriched textual data.

In some cases, we might opt to simplify the data representation using dimensionality reduction techniques [50]. This step helps in managing the complexity of high-dimensional embeddings [51] while retaining the essential features that influence clustering.With the semantic embeddings or reduced dimensions in hand, we apply our chosen clustering method. This algorithm groups documents that share similar themes, topics, or meanings. By identifying patterns in the embeddings, the algorithm effectively categorizes documents based on their semantic content.

To gauge the quality of our clusters, we employ appropriate evaluation metrics such as cluster purity, Normalized Mutual Information (NMI) [52], and Adjusted Rand Index (ARI) [53]. These metrics help us understand how well the clustering process has captured the inherent relationships within the data. By interpreting the clusters, we gain insights into the different themes or categories present in the documents. For enhanced understanding, visualization techniques can be employed to represent the clustered documents in a more easily interpretable format. This step can offer a clearer view of the patterns and groupings within the data.

The interpretation of clustering results involves analyzing the identified clusters to unravel the underlying thematic connections among the documents. This step enables us to extract meaningful insights and understand the primary topics or trends present in the dataset.

4.2 Evaluation Metrics

In the realm of deep text clustering and document classification, assessing the performance of the applied algorithms is crucial. Evaluation metrics [52 and 53] provide quantifiable measures to gauge the quality of the clusters generated. Here, we delve into three fundamental evaluation metrics: Cluster Purity, Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI).

4.2.1 Cluster Purity: It serves as a straightforward yet insightful metric to evaluate clustering effectiveness. It measures how well documents within a cluster truly belong to a specific category. The idea is to identify the majority class within each cluster and calculate the proportion of documents that share this common category. Higher cluster purity indicates that most documents in a cluster are accurately classified, contributing to the overall quality of the clustering outcome.

4.2.2 Normalized Mutual Information (NMI): NMI [52] measures the degree of similarity between the true labels of documents and the clusters assigned by the algorithm. It considers both precision and recall, capturing both the quantity of accurately grouped documents and the alignment of clusters with the actual categories. NMI values range from 0 to 1, where higher values indicate better alignment between the true categories and the generated clusters.

4.2.3 Adjusted Rand Index (ARI): The ARI [53] is a versatile metric that accounts for chance clustering and offers an adjusted evaluation. It measures the similarity between the true class assignments and the predicted clusters, factoring in the probability of accidental agreements. The ARI ranges from -1 to 1, where negative values signify chance-level clustering, 0 indicates random clustering, and higher positive values indicate stronger agreement between the true and predicted assignments.

In the context of our deep text clustering and document classification endeavor, these evaluation metrics provide a comprehensive understanding of how well the applied algorithms have performed. By employing these metrics, we can quantitatively assess the quality of the clusters, alignment with actual categories, and the accuracy of the clustering process.

V. COMPARATIVE ANALYSIS

In this section of the paper, we delve into a comprehensive comparative analysis to shed light on the performance and efficacy of deep text clustering algorithms [37] in the realm of document classification. Through a series of sub-sections, we explore the contrasts and implications of traditional clustering methods versus modern deep clustering techniques. We also investigate the influence of diverse embeddings on clustering accuracy, emphasizing the significance of appropriate semantic representations.

5.1 Traditional vs. Deep Text Clustering

In this sub-section, we conduct an in-depth exploration (as shown in table-1) of the performance disparities between traditional clustering techniques and the emerging domain of deep clustering.

Table-1 Comparative Analysis: Traditional vs. Deep Clustering Techniques for Document Classification

Clustering Methods	**Traditional Techniques**	**Deep Clustering Techniques**
Algorithm	K-means, Hierarchical Clustering, DBSCAN	CNNs, RNNs, Auto encoders, Self-Attention Mechanisms
Performance Metrics	Precision, Efficiency, Adaptability	Precision, Efficiency, Adaptability
Datasets	Various Textual Datasets	Various Textual Datasets
Evaluation Metrics	Cluster Purity, NMI, ARI	Cluster Purity, NMI, ARI
Insights	Comparing performance disparities	Highlighting enhanced accuracy and semantic understanding
Results	Statistical comparisons	Quantitative improvement in document classification
Observations	Traditional methods' limitations	Deep clustering's efficacy in capturing nuances
Implications	Traditional methods' relevance	Deep clustering's transformative impact
Conclusion	Traditional clustering vs. Deep clustering	Leveraging deep learning for superior outcomes

We meticulously compare the capabilities of methods such as K-means [25], hierarchical clustering [27], and DBSCAN [29] against CNNs [42], RNNs [35], Autoencoders [37], and Self-Attention Mechanisms [48] against 20 news group [54], AG News [55] and Reuters-21578 [56] datasets. By employing various datasets and evaluation metrics, we assess their precision, efficiency, and adaptability to the nuances of textual data is presented in table-2. This analysis provides insights into the enhanced accuracy and semantic understanding that deep clustering algorithms bring to document classification tasks.

Table-2 Comparative Performance of Clustering Methods on 20 News Group Dataset

Method	Dataset	Precision	Recall	F1-Score	NMI	ARI
K-means	20 News group	0.65	0.67	0.66	0.42	0.38
Hierarchical Clustering	20 News group	0.72	0.7	0.71	0.51	0.45
DBSCAN	20 News group	0.58	0.62	0.6	0.35	0.32
CNNs	20 News group	0.8	0.82	0.81	0.65	0.58
RNNs	20 News group	0.75	0.78	0.76	0.58	0.52
Autoencoders	20 News group	0.71	0.72	0.71	0.48	0.42
Self-Attention Mechanism	20 News group	0.82	0.85	0.83	0.7	0.62

Table-2 compares various clustering methods using multiple metrics on the "20 Newsgroups" dataset [54], providing insights into their performance and capabilities for text clustering and document classification. K-means achieves reasonable precision (0.65) and recall (0.67), but struggles with intricacies, evident from lower NMI and ARI. Hierarchical Clustering improves accuracy (precision 0.72, recall 0.70). DBSCAN faces challenges in complex data structures (precision 0.58, recall 0.62, lower NMI and ARI).

CNNs excel with high NMI (0.65) and ARI (0.58), capturing semantic relationships. RNNs perform well (NMI 0.58), capturing sequential dependencies. Autoencoders show moderate performance (precision 0.70, recall 0.72) while learning meaningful features. Self-Attention Mechanisms stand out (precision 0.82, recall 0.85), highlighting semantic understanding. High NMI (0.70) and ARI (0.62) underscore its proficiency in capturing text patterns.

The belowtable-3offers a concise comparison of diverse clustering techniques applied to the "Reuters-21578" dataset [56], illustrating their performance in the realm of text clustering and document classification.

Table-3 Comparative Performance of Clustering Methods on Reuters-21578Dataset

Method	Dataset	Precision	Recall	F1-Score	NMI	ARI
K-means	Reuters-21578	0.68	0.71	0.69	0.48	0.45
Hierarchical Clustering	Reuters-21578	0.76	0.78	0.77	0.58	0.54
DBSCAN	Reuters-21578	0.6	0.65	0.62	0.39	0.36
CNNs	Reuters-21578	0.82	0.83	0.82	0.68	0.64
RNNs	Reuters-21578	0.77	0.79	0.78	0.54	0.5
Autoencoders	Reuters-21578	0.71	0.73	0.72	0.45	0.41
Self-Attention Mechanism	Reuters-21578	0.85	0.86	0.85	0.72	0.68

K-means showcases a moderate performance with respectable precision (0.68) and recall (0.71), contributing to a balanced F1-Score (0.69), NMI (0.48), and ARI (0.45). Hierarchical Clustering yields improved results, demonstrating higher precision (0.76) and recall (0.78), resulting in a well-balanced F1-Score (0.77), NMI (0.58), and ARI (0.54).DBSCAN, although achieving reasonable precision (0.6) and recall (0.65), yields a comparatively lower F1-Score (0.62), NMI (0.39), and ARI (0.36). In contrast, CNNs exhibit strong capabilities, leading to high precision (0.82) and recall (0.83), thus reflecting an impressive F1-Score (0.82), NMI (0.68), and ARI (0.64).

RNNs show commendable results with precision (0.77) and recall (0.79), resulting in a balanced F1-Score (0.78), while Autoencoders demonstrate moderate performance with precision (0.71) and recall (0.73), resulting in a reasonable F1-Score (0.72).The Self-Attention Mechanism method stands out with high precision (0.85) and recall (0.86), leading to an impressive F1-Score (0.85), NMI (0.72), and ARI (0.68), emphasizing its potential in capturing semantic nuances and fostering accurate text clustering.

Table-4 offers a concise and informative comparison of different clustering methods applied to the "AG News" dataset [55], providing insights into their performance in the context of text clustering and document classification.

Table-4 Comparative Performance of Clustering Methods on AG News Dataset

Method	Dataset	Precision	Recall	F1-Score	NMI	ARI
K-means	AGNews	0.72	0.73	0.72	0.52	0.49
Hierarchical Clustering	AGNews	0.78	0.79	0.78	0.61	0.57
DBSCAN	AGNews	0.64	0.68	0.66	0.42	0.39
CNNs	AGNews	0.84	0.85	0.84	0.72	0.68
RNNs	AGNews	0.79	0.81	0.8	0.58	0.54
Autoencoders	AGNews	0.73	0.75	0.74	0.48	0.44
Self-Attention Mechanism	AGNews	0.87	0.88	0.87	0.76	0.72

K-means achieves moderate precision (0.72) and recall (0.73), contributing to a balanced F1-Score (0.72), NMI (0.52), and ARI (0.49). Hierarchical Clustering shows higher precision (0.78) and recall (0.79), resulting in a balanced F1-Score (0.78), NMI (0.61), and ARI (0.57). DBSCAN demonstrates moderate precision (0.64) and recall (0.68), leading to a reasonable F1-Score (0.66), NMI (0.42), and ARI (0.39).CNNs exhibit strong capabilities with high precision (0.84) and recall (0.85), resulting in an impressive F1-Score (0.84), NMI (0.72), and ARI (0.68). RNNs perform commendably with precision (0.79) and recall (0.81), leading to a balanced F1-Score (0.80), while Autoencoders show moderate precision (0.73) and recall (0.75), contributing to a reasonable F1-Score (0.74). The Self-Attention Mechanism approach stands out, showcasing high precision (0.87) and recall (0.88), resulting in an impressive F1-Score (0.87), NMI (0.76), and ARI (0.72). This underscores its proficiency in capturing intricate semantic relationships, enhancing accurate and meaningful text clustering.

The presented tables underscore the substantial power of deep text clustering in the realm of document classification. When compared to traditional methods, deep clustering techniques, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Autoencoders, and Self-Attention Mechanisms, consistently exhibit superior performance across multiple evaluation metrics and datasets.

Especially, CNNs and Self-Attention Mechanisms consistently stand out as top performers, showcasing their remarkable ability to capture semantic nuances, patterns, and relationships within textual data. These deep learning techniques consistently achieve high values in terms of precision, recall, F1-Score, Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). This emphasizes their effectiveness in uncovering meaningful clusters and accurately classifying documents, regardless of the dataset's nature or complexity.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

6.1 Interpretation of Results

The in-depth comparative analysis conducted across three distinct datasets provides a profound understanding of the performance of various deep text clustering methods in the realm of document classification. Remarkably, CNNs and Self-Attention Mechanisms emerge as frontrunners across all datasets, consistently surpassing the conventional techniques. This striking consistency underscores their resilience and effectiveness in capturing intricate semantic nuances and patterns inherent in textual data.

The elevated precision, recall, F1-Score, NMI, and ARI values exhibited by these techniques paint a vivid picture of their capability to not only cluster documents but also to comprehend the underlying semantic fabric. Their success signifies their capacity to transcend the limitations posed by traditional methods, delivering on the promise of more accurate and meaningful text clustering in the realm of document classification.

The robustness of CNNs and Self-Attention Mechanisms is particularly evident in their ability to glean meaningful insights from the complex and often unstructured nature of textual data. Their proficiency in capturing not only surface-level textual patterns but also the inherent semantic relationships within documents sets them apart as powerful tools for semantic understanding and accurate classification.

The consistent outperformance ofCNNs and Self-Attention Mechanisms over traditional methods across different datasets reaffirms their capability to elevate the accuracy and efficacy of document classification, ultimately enhancing our capacity to decipher and categorize the intricate layers of textual information.

6.2 Advantages and Challenges of Deep Text Clustering

6.2.1 Advantages:

Enhanced Semantic Understanding: Deep text clustering methods, notably CNNs and Self-Attention Mechanisms, stand out for their capability to capture intricate relationships and subtle nuances within textual data. This translates to a higher level of semantic understanding, allowing for more accurate and meaningful document classification.

Semantic Patterns and Relationships: Deep techniques excel in uncovering underlying semantic patterns and relationships that might be hidden in the intricate structure of textual content. This empowers them to recognize and group documents based on not just surface-level features but also deeper contextual meanings.

Superior Accuracy: The demonstrated superiority of CNNs and Self-Attention Mechanisms in terms of precision, recall, F1-Score, NMI, and ARI highlights their potential to provide more accurate clustering results compared to traditional methods. This accuracy is crucial for applications where precise categorization is vital, such as information retrieval or content recommendation.

Handling Unstructured Text: Deep methods outperform traditional techniques like K-means, hierarchical clustering, and DBSCAN in handling unstructured text data. They can effectively adapt to the complex and varied nature of textual content, mitigating challenges associated with diverse writing styles, languages, and formats.

6.2.2 Challenges:

Resource Intensive: Deep text clustering methods often demand more computational resources and time compared to traditional approaches. The complex neural architectures and large datasets require significant computational power for training and inference, making them less suitable for scenarios with limited resources.

Hyper parameter Tuning: Deep methods are sensitive to hyper parameter settings, which need to be carefully fine-tuned for optimal performance. Finding the right balance between model complexity, learning rates, and other parameters can be a challenging and time-consuming process.

Data Preprocessing: Despite their capabilities, deep methods can be sensitive to the quality and preprocessing of input data. Ensuring consistent tokenization, removing noise, and handling rare or out-of-vocabulary words are essential steps to prevent negative impacts on clustering results.

Interpretability: Deep methods can sometimes lack transparency in terms of how they arrive at their clustering decisions. This can be a drawback when interpretability and explainability are critical, such as in legal or medical domains where decisions need to be justified.

Over-fitting: Deep models, if not appropriately regularized, might be prone to overfitting, especially when dealing with smaller datasets. Balancing model complexity and regularization becomes crucial to prevent over-fitting and ensure generalizability.

6.3 Potential Applications and Future Directions

The compelling achievements showcased by deep text clustering methods illuminate a rich landscape of potential applications that span across various domains. These innovative techniques hold the key to revolutionizing fields such as content recommendation, sentiment analysis, topic modeling, and information retrieval. By harnessing their capability to unravel intricate semantic relationships within textual data, these methods can substantially elevate the quality and depth of insights extracted from diverse sources of textual information.

Looking ahead, the trajectory of research in deep text clustering unfolds promising future directions that could further amplify their impact. A significant avenue lies in refining these methods to gracefully accommodate larger and more complex datasets. This evolution would entail devising efficient strategies to handle the computational demands of extensive textual corpora, thereby making deep clustering methods more accessible and practical for real-world applications on grand scales.

Moreover, the horizon of exploration extends towards hybrid approaches that synergistically unite the strengths of both traditional and deep clustering methods. The integration of these two realms holds the potential to forge a powerful alliance, combining the interpretability and simplicity of traditional methods with the sophisticated semantic understanding and pattern recognition of deep techniques. Such amalgamations could potentially result in solutions that not only achieve higher accuracy but also offer enhanced scalability, offering a compelling path forward in addressing the evolving challenges of text clustering and document classification.

At glance, the diverse applications of deep text clustering methods hold significant promise in reshaping various aspects of information analysis and extraction. As the frontier of research advances, optimizing their efficacy for larger datasets and fostering synergistic collaborations with traditional methods emerges as a strategic thrust. These endeavors not only serve to unlock new dimensions of semantic understanding but also pave the way for more robust, efficient, and comprehensive solutions in the landscape of text analysis and knowledge extraction.

VII. CONCLUSION

The culmination of this investigation presents a robust conclusion that encapsulates the findings, contributions, implications, and the future trajectory of document classification through the lens of deep text clustering. The journey of comparative analysis across multiple datasets and deep text clustering techniques has unraveled compelling insights. The empirical evidence showcased that Convolutional Neural Networks (CNNs) and Self-Attention Mechanisms outshine traditional methods across various evaluation metrics. These advanced techniques exhibit a remarkable capability to capture the intricate semantics and relationships inherent within textual data. The tables, each tailored to specific datasets, underscore the consistent excellence of these deep methods in enhancing precision, recall, F1-Score, Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI).

The contributions made through this study are twofold. Firstly, it sheds light on the potency of deep text clustering methods in the realm of document classification. This study establishes the supremacy of CNNs and Self-Attention Mechanisms in deciphering nuanced patterns within text, underlining their potential to revolutionize the field. Secondly, this work emphasizes the significance of comparative analysis, providing researchers, practitioners, and stakeholders with an informed perspective on selecting the most suitable method for their specific requirements. The implications of these insights resonate across domains like content recommendation, sentiment analysis, topic modeling, and information retrieval.

Looking ahead, the horizon of document classification navigates towards an exciting intersection with deep text clustering. The innovative nature of these techniques, coupled with their capacity to bridge the gap between semantic understanding and machine learning, charts a promising trajectory. Future research directions encompass refining these methods to gracefully tackle the complexities of larger datasets while also exploring hybrid approaches that combine the strengths of both traditional and deep clustering methods. Such a path not only opens avenues for more precise and scalable solutions but also unlocks new dimensions of insights from textual data. This study underscores the importance of harnessing the power of semantics and patterns within text, propelling the realm of information analysis and knowledge extraction into an era of enhanced accuracy and understanding.

REFERENCES

[1] Lavanya, P. M., and E. Sasikala. "Deep learning techniques on text classification using Natural language processing (NLP) in social healthcare network: A comprehensive survey." In 2021 3rd international conference on signal processing and communication (ICPSC), pp. 603-609. IEEE, 2021.

[2] Guan, Renchu, Hao Zhang, Yanchun Liang, Fausto Giunchiglia, Lan Huang, and Xiaoyue Feng. "Deep feature-based text clustering and its explanation." IEEE Transactions on Knowledge and Data Engineering 34, no. 8 (2020): 3669-3680.

[3] Hassani, Hossein, Christina Beneki, Stephan Unger, Maedeh Taj Mazinani, and Mohammad Reza Yeganegi. "Text mining in big data analytics." Big Data and Cognitive Computing 4, no. 1 (2020): 1.

[4] Ezugwu, Absalom E., Abiodun M. Ikotun, and Andronicus A. Akinyelu. "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects." Engineering Applications of Artificial Intelligence 110 (2022): 104743.

[5] Maylawati, D. Saadillah, Tedi Priatna, Hamdan Sugilar, and M. Ali Ramdhani. "Data science for digital culture improvement in higher education using K-means clustering and text analytics." International Journal of Electrical and Computer Engineering 10, no. 5 (2020): 4569-4580.

[6] Ibrahim, R., S. Zeebaree, and K. Jacksi. "Survey on semantic similarity based on document clustering." Adv. sci. technol. eng. syst. j 4, no. 5 (2019): 115-122.

[7] Abualigah, Laith, Amir H. Gandomi, Mohamed Abd Elaziz, Husam Al Hamad, Mahmoud Omari, Mohammad Alshinwan, and Ahmad M. Khasawneh. "Advances in meta-heuristic optimization algorithms in big data text clustering." Electronics 10, no. 2 (2021): 101.

[8] Kokkinos, Konstantinos, and Eftihia Nathanail. "Exploring an ensemble of textual machine learning methodologies for traffic event detection and classification." Transport and Telecommunication 21, no. 4 (2020): 285-294.

[9] Buenano-Fernandez, Diego, Mario Gonzalez, David Gil, and Sergio Luján-Mora. "Text mining of open-ended questions in self-assessment of university teachers: An LDA topic modeling approach." Ieee Access 8 (2020): 35318-35330.

[10] Chen, Hongshu, Ximeng Wang, Shirui Pan, and Fei Xiong. "Identify topic relations in scientific literature using topic modeling." IEEE Transactions on Engineering Management 68, no. 5 (2019): 1232-1244.

[11] Tsapatsoulis, Nicolas, and Constantinos Djouvas. "Opinion mining from social media short texts: Does collective intelligence beat deep learning?." Frontiers in Robotics and AI 5 (2019): 138.

[12] Kumar, Yogesh, Komalpreet Kaur, and Gurpreet Singh. "Machine learning aspects and its applications towards different research areas." In 2020 International conference on computation, automation and knowledge management (ICCAKM), pp. 150-156. IEEE, 2020.

[13] Akpatsa, Samuel K., Xiaoyu Li, and Hang Lei. "A survey and future perspectives of hybrid deep learning models for text classification." In ICAIS 2021, Dublin, Ireland, July 19–23, pp. 358-369. Springer International Publishing, 2021.

[14] Yoon, Wonjin, Chan Ho So, Jinhyuk Lee, and Jaewoo Kang. "Collabonet: collaboration of deep neural networks for biomedical named entity recognition." BMC bioinformatics 20, no. 10 (2019): 55-65.

[15] R. Guan, et al.,"Deep Feature-Based Text Clustering and its Explanation" in IEEE Transactions on Knowledge & Data Engineering, vol. 34, no. 08, pp. 3669-3680, 2022.

[16] Pappagari, Raghavendra, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. "Hierarchical transformers for long document classification." In 2019 IEEE automatic speech recognition and understanding workshop (ASRU), pp. 838-844. IEEE, 2019.

[17] Kim, Sang-Woon, and Joon-Min Gil. "Research paper classification systems based on TF-IDF and LDA schemes." Human-centric Computing and Information Sciences 9 (2019): 1-21.

[18] Elnagar, Ashraf, Ridhwan Al-Debsi, and Omar Einea. "Arabic text classification using deep learning models." Information Processing & Management 57, no. 1 (2020): 102121.

[19] De Araujo, Pedro Henrique Luz, Teófilo Emídio de Campos, Fabricio Ataides Braz, and Nilton Correia da Silva. "VICTOR: a dataset for Brazilian legal documents classification." In Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 1449-1458. 2020.

[20] Chen, Liang, Shuo Xu, Lijun Zhu, Jing Zhang, Xiaoping Lei, and Guancan Yang. "A deep learning based method for extracting semantic information from patent documents." Scientometrics 125 (2020): 289-312.

[21] Kowsari, Kamran, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. "Text classification algorithms: A survey." Information 10, no. 4 (2019): 150.

[22] Rashid, Junaid, Muhammad Shafiq, and Akber Gardezi. "Topic modeling technique for text mining over biomedical text corpora through hybrid inverse documents frequency and fuzzy k-means clustering." IEEE Access 7 (2019): 146070-146080.

[23] Abasi, Ammar Kamal, Ahamad Tajudin Khader, Mohammed Azmi Al-Betar, Syibrah Naim, Zaid Abdi Alkareem Alyasseri, and Sharif Naser Makhadmeh. "A novel hybrid multi-verse optimizer with K-means for text documents clustering." Neural Computing and Applications 32 (2020): 17703-17729.

[24] Amer, Ali A., and Hassan I. Abdalla. "A set theory based similarity measure for text clustering and classification." Journal of Big Data 7 (2020): 1-43.

[25] Jo, Tae-Ho. "Inverted index based modified version of k-means algorithm for text clustering." Journal of Information Processing Systems 4, no. 2 (2008): 67-76.

[26] Venkataramanan, A. R. ., Kanimozhi, K. V. ., Valarmathia, K. ., Therasa, M. ., Hemalatha, S. ., Thangamani, M. ., & Gulati, K. . (2023). A Survey on Covid-19 & Its Impacts. International Journal of Intelligent Systems and Applications in Engineering, 11(3s), 129 –. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2550

[27] Pappagari, Raghavendra, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. "Hierarchical transformers for long document classification." In 2019 IEEE automatic speech recognition and understanding workshop (ASRU), pp. 838-844. IEEE, 2019.

[28] Stein, Roger Alan, Patricia A. Jaques, and Joao Francisco Valiati. "An analysis of hierarchical text classification using word embeddings." Information Sciences 471 (2019): 216-232.

[29] Meng, Yu, Jiaming Shen, Chao Zhang, and Jiawei Han. "Weakly-supervised hierarchical text classification." In Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, pp. 6826-6833. 2019.

[30] Cretulescu, Radu George, Daniel Morariu, Macarie Breazu, and Danie Volovici. "DBSCAN algorithm for document clustering." International Journal of Advanced Statistics and IT&C for Economics and Life Sciences 9, no. 1 (2019).

[31] Liu, Zhiwei, and Yan Yang. "Research on web text clustering based on DBSCAN optimization algorithm." In 6th International Workshop on Advanced Algorithms and Control Engineering (IWAACE 2022), vol. 12350, pp. 550-555. SPIE, 2022.

[32] Mohammed, Shapol M., Karwan Jacksi, and Subhi RM Zeebaree. "Glove word embedding and DBSCAN algorithms for semantic document clustering." In 2020 International Conference on Advanced Science and Engineering (ICOASE), pp. 1-6. IEEE, 2020.

[33] Widiastuti, N. I. "Convolution neural network for text mining and natural language processing." In IOP Conference Series: Materials Science and Engineering, vol. 662, no. 5, p. 052010. IOP Publishing, 2019.

[34] Aich, Satyabrata, Sabyasachi Chakraborty, and Hee-Cheol Kim. "Convolutional neural network-based model for web-based text classification." International Journal of Electrical & Computer Engineering (2088-8708) 9, no. 6 (2019).

[35] Wang, Fan, Jing-Fang Yang, Meng-Yao Wang, Chen-Yang Jia, Xing-Xing Shi, Ge-Fei Hao, and Guang-Fu Yang. "Graph attention convolutional neural network model for chemical poisoning of honey bees' prediction." Science Bulletin 65, no. 14 (2020): 1184-1191.

[36] Skrlj, Blaz, Jan Kralj, Nada Lavrac, and Senja Pollak. "Towards robust text classification with semantics-aware recurrent neural architecture." Machine Learning and Knowledge Extraction 1, no. 2 (2019): 34.

[37] Murthy, G. S. N., Shanmukha Rao Allu, Bhargavi Andhavarapu, Mounika Bagadi, and Mounika Belusonti. "Text based sentiment analysis using LSTM." Int. J. Eng. Res. Tech. Res 9, no. 05 (2020).

[38] Hosseini, Soodeh, and Zahra Asghari Varzaneh. "Deep text clustering using stacked AutoEncoder." Multimedia Tools and Applications 81, no. 8 (2022): 10861-10881.

[39] Yin, Hui, Xiangyu Song, Shuiqiao Yang, Guangyan Huang, and Jianxin Li. "Representation learning for short text clustering." 22nd International Conference on Web Information Systems Engineering, WISE 2021, Melbourne, VIC, Australia, October 26–29, 2021, 321-335. Springer International Publishing, 2021.

[40] Yilmaz, Seyhmus, and Sinan Toklu. "A deep learning analysis on question classification task using Word2vec representations." Neural Computing and Applications 32 (2020): 2909-2928.

[41] Gundogan, Esra, and Mehmet Kaya. "Research paper classification based on Word2vec and community discovery." In 2020 international conference on decision aid sciences and application (DASA), pp. 1032-1036. IEEE, 2020.

[42] Chen, Kai, Rabea Jamil Mahfoud, Yonghui Sun, Dongliang Nan, Kaike Wang, Hassan Haes Alhelou, and Pierluigi Siano. "Defect texts mining of secondary device in smart substation with GloVe and attention-based bidirectional LSTM." Energies 13, no. 17 (2020): 4522.

[43] Hossain, Md Rajib, and Mohammed Moshiul Hoque. "Covtexminer: Covid text mining using cnn with domain-specific glove embedding." In International Conference on Intelligent Computing & Optimization, pp. 65-74. Cham: Springer International Publishing, 2022.

[44] Mark White, Thomas Wood, Maria Hernandez, María González , María Fernández. Enhancing Learning Analytics with Machine Learning Techniques. Kuwait Journal of Machine Learning, 2(2). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/184

[45] Selva Birunda, S., and R. Kanniga Devi. "A review on word embedding techniques for text classification." Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020 (2021): 267-281.

[46] Talebpour, Mozhgan, Alba García Seco de Herrera, and Shoaib Jameel. "Topics in Contextualised Attention Embeddings." In European Conference on Information Retrieval, pp. 221-238. Cham: Springer Nature Switzerland, 2023.

[47] Alaparthi, Shivaji, and Manit Mishra. "BERT: A sentiment analysis odyssey." Journal of Marketing Analytics 9, no. 2 (2021): 118-126.

[48] Ali, Sikandar, Anam Nasir, Ali Samad, Samad Basser, and Azeem Irshad. "An automated approach for the prediction of the severity level of bug reports using GPT-2." Security and Communication Networks 2022 (2022).

[49] Soydaner, Derya. "Attention mechanism in neural networks: where it comes and where it goes." Neural Computing and Applications 34, no. 16 (2022): 13371-13385.

[50] Li, Weijiang, Fang Qi, Ming Tang, and Zhengtao Yu. "Bidirectional LSTM with self-attention mechanism and multi-

channel features for sentiment classification." Neurocomputing 387 (2020): 63-77.

[51] Kadhim, Ammar Ismael, Yu-N. Cheah, and Nurul Hashimah Ahamed. "Text document preprocessing and dimension reduction techniques for text document clustering." In 2014 4th international conference on artificial intelligence with applications in engineering and technology, pp. 69-73. IEEE, 2014.

[52] Huang, Xuan, Lei Wu, and Yinsong Ye. "A review on dimensionality reduction techniques." International Journal of Pattern Recognition and Artificial Intelligence 33, no. 10 (2019): 1950017.

[53] Harel, David, and Yehuda Koren. "Graph drawing by high-dimensional embedding." In International symposium on graph drawing, pp. 207-219. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002.

[54] Knops, Zeger F., JB Antoine Maintz, Max A. Viergever, and Josien PW Pluim. "Normalized mutual information based registration using k-means clustering and shading correction." Medical image analysis 10, no. 3 (2006): 432-439.

[55] Steinley, Douglas, Michael J. Brusco, and Lawrence Hubert. "The variance of the adjusted Rand index." Psychological methods 21, no. 2 (2016): 261.

[56] Borkar, Karishma, and Nutan Dhande. "Efficient text classification of 20 newsgroup dataset using classification algorithm." Int J Recent Innov Trends Comput Commun 5, no. 6 (2017): 1236-1240.

[57] Zhang, Xiang, Junbo Zhao, and Yann LeCun. "Character-level convolutional networks for text classification." Advances in neural information processing systems 28 (2015).

[58] Rodríguez, Juan M., Hernán D. Merlino, Patricia Pesado, and Ramón García-Martínez. "Evaluation of open information extraction methods using Reuters-21578 database." In Proceedings of the 2nd International Conference on Machine Learning and Soft Computing, pp. 87-92. 2018.