NeuroClustr: Empowering Biomedical Text Clustering with Neuro-Symbolic Intelligence

Fakhare Alam[1], Khalid Mahmood Malik[1], Madan Krishnamurthy
[1]Department of Computer Science & Engineering, Oakland University, Rochester, United States
{fakharealam, mahmood }@oakland.edu, madank1249@gmail.com

Abstract

The exponential growth of biomedical unstructured data being generated every day has made it increasingly challenging to accurately cluster them and extract meaningful insights. Traditional clustering algorithms such as K-means are limited in their ability to capture contextual information and semantically explain the reasoning behind the clustering results when applied to a large corpus of unstructured data. As a result, there is a need for more advanced techniques that can integrate deep learning and symbolic reasoning to improve clustering performance. Integrating domain-specific knowledge from external sources through a Neuro-Symbolic approach can facilitate the optimization of clustering algorithms by generating new hypotheses. This research paper proposes a novel framework NeuroClustr to cluster biomedical text corpus using a Neuro-Symbolic approach in conjunction with deep learning. The framework employs Recurrent Neural Network (RNN) architecture to capture important sequence to sequence information in textual data and uses BioBERT based encoded representation and infused knowledge rules from external sources such as domain specific ontology to effectively cluster the biomedical documents. The evaluation results show that the proposed framework outperforms traditional baseline models by 43% and achieves average precision of 88% across all identified clusters for COVID-19 Dataset. This demonstrates the potential of deep neural networks with knowledge infusion in improving clustering accuracy for large and complex biomedical text corpus.

1 Introduction

Biomedical research produces a vast amount of textual and unstructured data, including scientific publications, electronic health records, clinical trial reports, and patient forums. In the PubMed database alone, there are more than a million research articles published every year [Landhuis,2016]. Analyzing and curating information from this huge corpus of data is a complex task and presents a significant challenge in extracting information, measuring data and information qualities such as accuracy, consistency, completeness, timeliness, availability and scalability[Wilkinson, *et al*., 2016].Clustering techniques can be used to organize and group this large, complex, and heterogeneous textual data based on the contents and structure, providing insights and facilitating knowledge curation by developing domain specific ontologies and easing the knowledge discovery process. Clustering plays a crucial role in enhancing the effectiveness of information and knowledge management systems. It enables researchers and physicians to efficiently identify relevant publications, clinical trials, and patient forums related to their research questions, thereby saving time and effort in searching and reviewing large volumes of text. Furthermore, clustering can be used to identify emerging topics and trends in biomedical research, which can help to guide future research directions and identify potential collaborations. By identifying patterns and relationships between documents, clustering algorithms can reveal insights and patterns that might not be immediately apparent from a single document. In the biomedical domain, clustering can also be used to identify and group similar biomedical terms and concepts, enable entity recognition and relation extraction, which are critical components of many Natural Language Processing (NLP) applications [Zhang *et al*., 2013].

Clustering techniques are a key component of many machine learning algorithms. However, choosing the right clustering algorithm and evaluating the quality of the resulting clusters can be challenging tasks, as it depends on the specific characteristics of the dataset and domain specific knowledge. There are several clustering techniques used in the biomedical domain, including hierarchical clustering [Fionn *et al*.,2012], K-means clustering [Hartigan *et al*.,1979], spectral clustering [Von Luxburg,2007], and DBSCAN [Ester *et al.,1996*]. Although these algorithms are useful, they have several limitations. For instance, they can be sensitive to noise and outliers, have difficulty handling varying cluster size and density, and

incur performance overhead when working with large datasets. Moreover, these techniques lack context and have limitations in providing explanations for cluster results, thus creating a need to develop a Neuro-Symbolic clustering framework using deep learning and symbolic reasoning having an ability to process big data, capture contextual information and infuse knowledge from domain specific external sources to enable reasoning in the cluster identification, creation, and optimization.

By combining neural network models with symbolic representations, Neuro-Symbolic clustering can effectively capture the complex relationships and interactions between biomedical concepts and terms, even in the presence of noisy and ambiguous data. The ability for external knowledge infusion, reasoning, and explainability in clustering techniques are critical to improve the accuracy and quality of clustering results and enable actionable insights. Explainability and reasoning is crucial in ensuring that the clustering results are transparent and interpretable and follow FAIR principles [Wilkinson, *et al*., 2016], especially in domains such as healthcare, where the decisions made based on clustering results can have significant consequences.

This paper attempts to solve the limitations of the traditional clustering approach by proposing a deep neural network-based clustering framework-NeuroClustr with external knowledge infusion and reasoning capability in forms of rules to optimize and improve the cluster quality. **Fig.1.** describes the conceptual framework of the proposed architecture. The architecture contains four components namely Document Preprocessing for initial text cleaning, Training Data Generation module to create labeled information, Cluster Modeling to create and experiment with different algorithms, and External Knowledge Infusion to infuse domain specific information in the form of rules.

The main contribution of this paper can be summarized as follows:

- First, a state-of-the-art deep learning-based clustering component to capture contextual information using sequence to sequence architecture of Bidirectional Long Short-Term Memory (Bi-LSTM) and BioBERT embeddings.
- Second, an external knowledge infusion module to incorporate additional information in the form of rules by mining domain specific information leading to creation of new hypotheses and cluster optimization.

Lastly, we present a scalable clustering framework NeuroClustr to mine biomedical text data and create optimized clusters by generating new hypotheses utilizing knowledge from external sources.

The remainder of the paper is structured as follows. Section 2 reviews previous literature on clustering, its application in the biomedical domain, clustering construction methods. Section 3, material & methods provides the details about the dataset, the framework of deep clustering with Neuro-Symbolic, followed by results & evaluation in section 4. Section 5, Section 6, and Section 7 highlights contributions, limitations, and future research directions of this paper respectively.

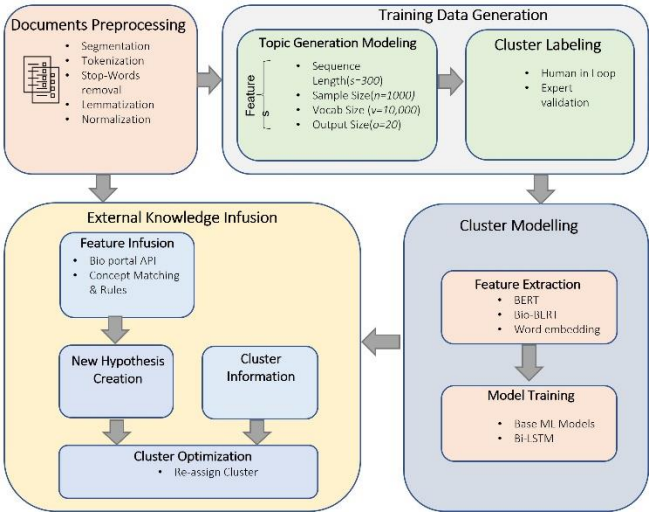

Fig. 1: Conceptual View of proposed neuro-symbolic deep clustering framework - NeuroClustr

2 Related Works

Deep clustering has several advantages over traditional methods where it can categorize unstructured, high-dimensional data by identifying complex patterns, implement dimensionality reduction and accomplish cluster assignments simultaneously to provide a highly scalable end-to-end framework. However, challenges such as quality of clustering based on document content/meaning; and categorizing documents with multiple topics [Anastasiu, et.al., 2018] must be evaluated. Language modeling techniques such as vector space modeling, dimensionality reduction, topic modeling and continuous space modeling use features to represent documents in a collection. Segmenting documents into sections of different topics can further enhance deep clustering. Authors in [Fard, *et al*.,2020] use constrained clustering in the form of seed words, while in the current study external knowledge infusion is employed using a neural network, facilitated by expert-labeled training data.

The use of deep clustering models in the big text corpus has been sought out and implemented as a significant tool for document analysis [Anastasiu, *et.al*., 2018]. Authors in [Karim *et al*.,2021] present the application of deep learning clustering algorithms for unsupervised research on three different use cases of bioimaging, cancer genomics and biomedical

text mining. They deploy a variant of recurrent neural network algorithm LSTM (long short-term memory networks) to achieve document clustering. Authors in [Kozawa et al., 2018, Jadhav et al.,2022] conducted an extensive analysis of the human body by employing body-wide modeling techniques. They developed advanced computational models using a vast biomedical dataset and harnessed the power of word embedding and text mining, employing ontology-based clustering methodologies. However, this approach exhibits a limitation in its capacity to capture contextual information effectively. This limitation arises since features are solely extracted based on the similarity between tokens, without considering the broader context in which these tokens appear. Work in [Davagdorj *et.al*., 2022] represents a biomedical document clustering framework based on BioBERT which is a pre-trained language representation technique. Either method has their own advantages in improving the deep clustering analysis, LSTM can better handle long-term dependencies and can model complex sequential data; BioBERT can capture multi-directional context with a faster training time [Gabralla *et al*.,2012]. Our current work leverages on the advantages of both methods and implements a deep clustering framework that involves LSTM and BioBERT to achieve better clustering accuracy.

Existing deep learning methods can be better enhanced by domain and conceptual knowledge infusion [Sheth et.al., 2019]. Knowledge infusion can co-use symbolic AI with data-driven AI to provide a class of neuro-symbolic AI methods called knowledge-infused learning (KiL) [Gaur *et al*.,2022]. Neuro-symbolic approaches can enhance the efficiency of clustering framework by including symbolic reasoning with neural perception. Work [Aspis, et.al., 2022] illustrates the generation of clusters using trained perceptions which are further labeled using symbolic knowledge. Researchers in [Venugopal et.al., 2021] overcome the drawbacks of deep neural networks such as long convergence times and overfitting data by taking the neuro-symbolic approach on big data. Big data is first converted into a symbolic model, followed by embedding to create a training set. Other works such as [Kursuncu *et al*., 2019] have explored KiL in deep learning models where infusion of representational external knowledge from knowledge graphs will aid in supervising the learning of features and enhance the model learning process. In our current neuro-symbolic deep clustering framework we implement the knowledge infusion module that infuses external knowledge into the penultimate layer of the trained model from domain specific ontologies for identified topics to optimize the cluster. This novel approach provides enhanced model performance as depicted in the results.

3 Material and Methods

3.1 Datasets and Preprocessing

Within the current clustering framework, we use COVID-19 Open Research Dataset (CORD-19) [Wang et.al., 2020] to compile research papers that pertain to COVID-19 and contain pertinent information regarding the virus. The CORD-19 database houses over one million academic articles on COVID-19, SARS-CoV-2, and other coronaviruses

3.2 NeuroClustr: Neuro-Symbolic Deep Clustering Framework

The proposed clustering framework comprises four key components: Data Preprocessing, Training Data Generation, Cluster Model, and Knowledge Infusion. The main tasks of the data processing module are cleaning, stop words removal, tokenization, and generating vector representations. The training data generation involves topic modeling utilizing word frequency approach and generating labeled dataset with expert input. The model training module includes creating sequence to sequence models based on the labeled data. Lastly, the knowledge infusion module takes the penultimate layer output and optimizes the cluster initially generated by k-means cluster by infusion of external knowledge from specific ontologies for identified topics. **Fig.2.** depicts the distinct constituents of the proposed clustering framework.

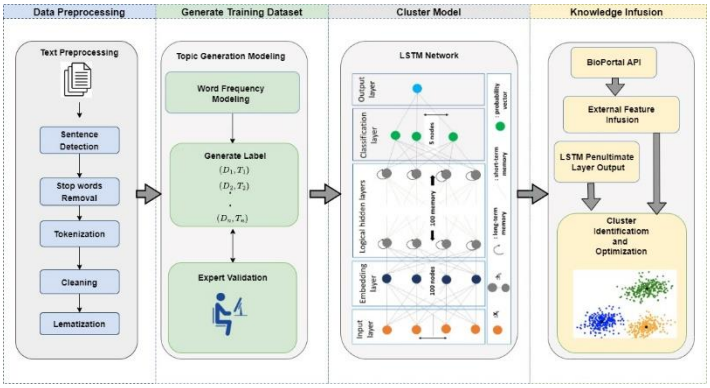

Fig. 2: Functional View of Neuro-Symbolic deep clustering framework - NeuroClustr

Data Preprocessing

The COVID-19 dataset consists of research articles from diverse data sources. Data preprocessing involves eliminating inconsistencies and duplicates in the data, performing text cleaning for NLP, and converting the text into a vectorized format that can be embedded and utilized by machine learning models.

Text Cleaning

In the text cleaning process, several techniques were employed to process the full-text research articles. These techniques included the removal of special characters, HTML tags, references, tables, and images. Following this, sentences

were identified and punctuation and stop words were removed from the text using the Python Natural Language Toolkit (NLTK) library [Loper *et.al*., 2002], which was augmented with medical-specific stop words. To identify special characters, symbols, and URLs, regular expressions were utilized. Once these text cleaning steps were completed, the final output was deemed to be cleaned and ready for tokenization.

Tokenization

The next step in our methodology involved the tokenization of sentences from the research articles, thereby dividing them into individual tokens. In order to extract concepts from the texts, we relied on the NLTK library to generate relevant concepts from the research papers. To facilitate the use of machine learning models, we leveraged an in-built tokenizer that was available in large language models (LLMs) such as BERT and BioBERT and generated an embedding matrix.

3.2 Generate Training Dataset

In this module, we leveraged the concepts generated from each document to generate word frequency metrics for every research article. By mapping the concepts to the articles, we were able to identify various types of documents, such as those related to drugs, vaccines, symptoms, and genetics.

To refine the word frequency metrics further, we selected the top 100 most frequently occurring concepts from each article. This provided us with a concise and informative representation of the article's content. Next, we utilized expert evaluation to label the data into four distinct categories: drugs, vaccines, symptoms, and genetics. By doing so, we were able to transform the unstructured text data island and create labeled data.

This approach allowed us to generate a labeled dataset that could be used for training machine learning models to classify research articles based on their content. The resulting dataset, enriched with expert labeling, provided a more accurate and targeted approach for analyzing the vast amount of research on COVID-19 available in the literature.

3.3 Cluster Model

The aim of this model is to develop a clustering model, which can understand the textual context and classify the document into four identified categories: Drug, Vaccine, Symptom and Genetics. Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that has shown great success in sequence modeling tasks such as natural language processing. In this experiment, we used BioBERT Embeddings [Lee, *et al*., 2020] to vectorize the text data. BioBERT is a domain-specific language model that was pre-trained on a large corpus of biomedical text. BioBERT has been shown to outperform other general-purpose language models on a variety of biomedical natural language processing tasks.

The LSTM architecture consists of five main layers: the input layer, the embedding layer, the hidden layer, the classification layer, and the output layer. The embedding layer is responsible for transforming the input sequence into a 100-dimensional representation that captures the semantic meaning of each document. The hidden layer utilizes 100 memory units and a recurrent dropout rate of 0.2 to capture temporal dependencies in the sequence. The output layer is a dense layer that employs a SoftMax activation function to produce four nodes, corresponding to the four categories for which the model is trained. The categorical cross-entropy loss function is utilized to optimize the model's performance during training.

To train the model, we utilized a batch size of 64 and trained the model for 5 epochs. During training, we randomly selected 100 documents to serve as a validation set, which was used to monitor the model's performance.

3.4 Knowledge Infusion

The domain specific knowledge using external data sources can enable explanation, reasoning by creating new hypotheses and optimizing clustering. Biomedical repositories such as those in the field of healthcare and life sciences contain valuable information that is relevant to the identified clusters. In this experiment we used BioBERT Embeddings with above mentioned LSTM architecture and mined topic specific information using domain specific ontologies using BioPortal API [Noy *et.al*., 2009]. We followed the three steps process to perform knowledge infusion. At first, we extracted the concept matched to specific ontologies for each research article and calculated the normalized portion of concept matched to total concepts per research article. In the next step, we fetched synonyms, prefLabel for these concepts and created the embedding using BioBERT. Finally, we combined these features with LSTM penultimate layer output and infused this additional knowledge. The embedding of this additional knowledge created a new hypothesis on top of the clustering done using LSTM penultimate layer output and optimized the output of initial clustering results. **Table 1** presents the specific ontologies utilized for extracting knowledge and creating features, along with their respective types and names. **Algorithm 1** provides a comprehensive overview of the entire knowledge infusion process, outlining the step-by-step procedure for incorporating the additional knowledge into the system.

Cluster	Ontologies
Drug (D)	Drug Ontology (DRON)
	Prescription of Drugs Ontology (PDRO)
Vaccines (V)	Vaccine Ontology (VO)
	Vaccine Investigation Ontology (VIO)
	Vaccine Informed Consent Ontology (VICO)
Symptom (S)	Symptom Ontology (SYMP)

	Clinical Signs and Symptom Ontology (CSSO)
Genetic (G)	Gene Ontology
	Gene Ontology Extension

Table1: Ontologies for external knowledge infusion

Algorithm 1 Knowledge Infusion Algorithm

Input: Documents [$d_1 \ldots d_{1000}$], LSTM penultimate layer output [$d_1 \ldots d_{1000}$]
Output: Optimized Cluster [1,2,3,4]
1: Let d= 1.
2: **while** d<=1000 **do**
while ontology in *[Drug, Vaccine. Symptom, Genetics]*
3: Fetch concepts from ontology
4: Calculate concept match
5: Normalize concept percent match
6: **Concept Match Pct**= Normalized concept match
7: Fetch Synonyms, PrefLabel from ontologies
8: Create semantic embedding using BioBERT
9: **Semantic Embedding**=BioBERT Embedding
10: Concatenate Features = [LSTM penultimate layer, Concept Match pct, Semantic Embedding]
11: **end while**
12: **end while**
13: **return** Concatenated Features

4 Results and Evaluations

4.1 Dataset Preparation

To evaluate the proposed clustering framework, a randomized sample of 1000 research articles from the CORD-19 dataset was selected and labeled using topic frequency modeling and human effort.

4.2 Experimental Setup

To assess the quality of the clustering outcome, we conducted experiments using machine learning models, including K-means, Zero Shot Learning (ZSL) [Xian et.al., 2017] and RNNs such as LSTM. Additionally, we utilized both generic embedding techniques like BERT and domain-specific embeddings such as BioBERT. These experiments allowed us to evaluate the effectiveness of different approaches and determine the most suitable techniques for our specific domain.

K-Means Clustering
K-means clustering is a widely used unsupervised machine learning algorithm that partitions data into K clusters based on similarities in the data. One of the primary advantages of k-means clustering is its scalability to large datasets. Additionally, it is a simple and fast algorithm that can be applied to a wide range of applications such as image segmentation, customer segmentation, and anomaly detection. At first, we used BERT encoding to generate vectorized input, followed by PCA [Abdi, *et al*., 2010] to reduce the dimension based on variance explanation. Next, we used elbow method [Kodinariya *et al*., 2013] to get the initial number of clusters (*n=4*). The dimensionally reduced vector and optimized number of clusters are fed to the K-means model to generate the cluster and later it is mapped back to original documents to specify the name of the clusters.

Zero Shot Learning with BERT embeddings
BERT (Bidirectional Encoder Representations from Transformers) [Delvin et.al., 2018] is a pre-trained language model that has achieved state-of-the-art results on various NLP tasks. Zero-Shot learning (ZSL) is a type of machine learning technique that enables models to recognize and classify objects or concepts that have not been seen during training. This approach allows the model to generalize to new categories by learning a mapping between different domains. One of the key advantages of zero-shot learning is its ability to overcome the limitations of supervised learning, where the model is trained only on labeled data. With zero-shot learning, models can be trained on a smaller set of labeled data and still generalize to new categories. We used ZSL by using transformer architecture and embeddings generated by BERT tokenizer.

LSTM with BERT, BioBERT and Knowledge Infusion
In this approach, BERT and BioBERT embeddings are used to capture contextual information from the input text, which is then fed into an LSTM layer to capture the sequence information. The LSTM layer can then model the long-term dependencies in the sequence and make predictions based on the contextual and sequential information learned from the embeddings. We used the same LSTM architecture as presented in methods. In the next iteration, we extracted knowledge in from domain specific ontologies and infused it with the clustering result to create new hypotheses and optimize the cluster output.

4.3 Cluster Model Results

Table 2 presents a comparative analysis of model performance across different categories using precision as evaluation metric. The results indicate that k-means clustering with BERT-encoded vectors had an average precision of 45%. Zero-shot learning with BERT vectors improved the precision by 22% due to the transformer architecture of ZSL, which can understand contextual information and hold important sequences using its attention layer. The LSTM model with BERT encoding further improved the performance as it was trained on the labeled dataset, unlike k-means clustering and ZSL, which had no labeled dataset. The LSTM model with BioBERT embeddings performed the best with an average precision of 76%, an improvement of 5% on the LSTM model with BERT. The specificity of embeddings in the biomedical domain is the reason behind its effectiveness, as it provides a deep contextualization of vector inputs with domain-specific concepts.

Finally, the best-performing model was the LSTM with BioBERT embedding with external knowledge infusion using specific biomedical ontologies, which had an average precision of 88%. This indicates that the external knowledge was able to aid in the existing clustering mechanism and improved the performance by 12%. **Table 3** presents a detailed performance evaluation using precision, recall, and F1-score.

Topics / ML Model	D	V	S	G	Avg
K-means Clustering BERT	0.47	0.42	0.53	0.37	0.45
ZSL BERT	0.69	0.63	0.67	0.68	0.67
LSTM BERT	0.72	0.74	0.62	0.77	0.71
LSTM BioBERT	0.73	0.69	0.79	0.82	0.76
LSTM BioBERT Knowledge Infusion	0.89	0.82	0.91	0.90	0.88

Table 2: Comparative result (Precision) of traditional model and LSTM with Knowledge Infusion across categories Drug(D), Vaccine(V), Symptom(S), Genetics(S)

Topics	Precision	Recall	F1-Score
Drug	0.89	0.81	0.84
Vaccines	0.82	0.81	0.81
Symptoms	0.91	0.87	0.84
Genetic	0.90	0.86	0.87

Table 3: Precision, Recall, F1-Score of proposed LSTM model with external knowledge infusion

5 Discussion

Neuro-symbolic approaches combine the strengths of both symbolic and neural approaches to enable reasoning and explanation over structured and unstructured data [Yu et.al., 2021]. In clustering, this approach can help to improve the quality of clustering by enabling the incorporation of domain-specific knowledge and prior information in the clustering process as visible by the significant improvement in the performance of the model shown in **Table 3.**This modeling approach addresses the challenges in clustering such as handling high-dimensional data and overcoming the limitations of traditional unsupervised clustering algorithms leading to improved performance, enhanced interpretability, and more efficient knowledge discovery in various domains including healthcare and biomedicine.

Finally, the use of large language model embedding such as BERT and BioBERT in conjunction with LSTM models and external knowledge infusion outperforms the traditional k-means clustering algorithm and has 40% more precision in clustering documents.

5 Limitation and Future Directions

The proposed clustering framework has been evaluated on the biomedical domain, demonstrating its adaptability to other domains. Future research will aim to enhance the methodology and assess its applicability to other domains such as legal and financial documents for extracting key topics and generating insights. To overcome the need for large amounts of labeled data, the LSTM network could be further developed using few-shot learning experiments. Additionally, incorporating semantic information into the existing concept matching features from external sources could further improve the overall framework's performance. This will enable the framework to better understand the context of the documents and enhance the accuracy of the clustering results.

6 Conclusion

The neuro-symbolic clustering framework presented in this research article is a promising approach for document clustering. The integration of deep learning and symbolic reasoning techniques has shown to improve the clustering performance significantly, particularly in domains such as biomedical research. The experiments conducted in this study demonstrate the effectiveness of the proposed framework in achieving high precision. Furthermore, the framework is flexible enough to be applied to various domains and can be extended in multiple ways to improve its performance. Overall, the results indicate that the neuro-symbolic approach has the potential to significantly enhance the efficiency and accuracy of document clustering, thereby aiding in knowledge discovery and decision-making processes in various fields.

Ethical Statement

There are no ethical issues.

References

[Abdi *et al*., 2010] Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." *Wiley interdisciplinary reviews: computational statistics* 2, no. 4 (2010): 433-459.

[Anastasiu, *et.al*., 2018] Anastasiu, David C., Andrea Tagarelli, and George Karypis. "Document clustering: the next frontier." In *Data Clustering*, pp. 305-338. Chapman and Hall/CRC, 2018.

[Aspis, *et.al*., 2022] Aspis, Yaniv, Krysia Broda, Jorge Lobo, and Alessandra Russo. "Embed2Sym-Scalable Neuro-Symbolic Reasoning via Clustered Embeddings." In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, vol. 19, no. 1, pp. 421-431. 2022.

[Davagdorj *et.al*., 2022] Davagdorj, Khishigsuren, Kwang Ho Park, Tsatsral Amarbayasgalan, Lkhagvadorj Munkhdalai, Ling Wang, Meijing Li, and Keun Ho Ryu. "Biobert based efficient clustering framework for biomedical document analysis." In *Genetic and Evolutionary Computing: Proceedings of the Fourteenth International Conference on Genetic and Evolutionary Computing, October 21-23, 2021, Jilin, China*, pp. 179-188. Singapore: Springer Nature Singapore, 2022.

[Delvin *et.al*., 2018] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[Ester *et al.,1996*] Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." In *kdd*, vol. 96, no. 34, pp. 226-231. 1996.

[Fard *et al*.,2020] Fard, Mazar Moradi, Thibaut Thonet, and Eric Gaussier. "Seed-guided deep document clustering." In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42*, pp. 3-16. Springer International Publishing, 2020.

[Fionn *et al*.,2012] Murtagh, Fionn, and Pedro Contreras. "Algorithms for hierarchical clustering: an overview." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2, no. 1 (2012): 86-97.

[Gaur *et al*.,2022] Gaur, Manas, Kalpa Gunaratna, Shreyansh Bhatt, and Amit Sheth. "Knowledge-Infused Learning: A Sweet Spot in Neuro-Symbolic AI." *IEEE Internet Computing* 26, no. 4 (2022): 5-11.

[Gabralla *et al*.,2012] Gabralla, Lubna, and Haruna Chiroma. "Deep learning for document clustering: a survey, taxonomy and research trend." *Journal of Theoretical and Applied Information Technology* 98, no. 22 (2020): 3602-3634.

[Hartigan *et al*.,1979] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the royal statistical society. series c (applied statistics), 28(1), 100-108.

[Jadhav *et al*.,2022] Jadhav, Aditya, Tarun Kumar, Mohit Raghavendra, Tamizhini Loganathan, and Manikandan Narayanan. "Predicting cross-tissue hormone–gene relations using balanced word embeddings." Bioinformatics 38, no. 20 (2022): 4771-4781.

[Karim *et al*.,2021] Karim, Md Rezaul, Oya Beyan, Achille Zappa, Ivan G. Costa, Dietrich Rebholz-Schuhmann, Michael Cochez, and Stefan Decker. "Deep learning-based clustering approaches for bioinformatics." *Briefings in Bioinformatics* 22, no. 1 (2021): 393-415.

[Kodinariya *et al*., 2013] Kodinariya, Trupti M., and Prashant R. Makwana. "Review on determining number of Cluster in K-Means Clustering." *International Journal* 1, no. 6 (2013): 90-95.

[Kozawa et al., 2018] Kozawa, Satoshi, Ryosuke Ueda, Kyoji Urayama, Fumihiko Sagawa, Satsuki Endo, Kazuhiro Shiizaki, Hiroshi Kurosu et al. "The body-wide transcriptome landscape of disease models." Iscience 2 (2018): 238-268.

[Kursuncu *et al*., 2019] Kursuncu, Ugur, Manas Gaur, and Amit Sheth. "Knowledge infused learning (k-il): Towards deep incorporation of knowledge in deep learning." *arXiv preprint arXiv:1912.00512* (2019).

[Landhuis, 2016] Landhuis, Esther. "Scientific literature: Information overload." *Nature* 535, no. 7612 (2016): 457-458.

[Lee *et al*., 2020] Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics* 36, no. 4 (2020): 1234-1240.

[Loper *et.al*., 2002] Loper, Edward, and Steven Bird. "Nltk: The natural language toolkit." arXiv preprint cs/0205028 (2002).

[Noy *et.al*., 2009] Noy, Natalya F., Nigam H. Shah, Patricia L. Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet et al. "BioPortal: ontologies and integrated data resources at the click of a mouse." *Nucleic acids research* 37, no. suppl_2 (2009): W170-W173.

[Sheth *et.al*., 2019] Sheth, Amit, Manas Gaur, Ugur Kursuncu, and Ruwan Wickramarachchi. "Shades of knowledge-infused learning for enhancing deep learning." *IEEE Internet Computing* 23, no. 6 (2019): 54-63.

[Venugopal *et.al*., 2021] Venugopal, Deepak, Vasile Rus, and Anup Shakya. "Neuro-Symbolic Models: A Scalable, Explainable Framework for Strategy Discovery from Big Edu-Data." In *Proceedings of the 2nd Learner Data Institute Workshop in Conjunction with The 14th International Educational Data Mining Conference*. 2021.

[Von Luxburg, 2007] Von Luxburg, Ulrike. "A tutorial on spectral clustering." *Statistics and computing* 17 (2007): 395-416

[Wang *et.al*., 2020] Wang, Lucy Lu, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk et al. "Cord-19: The covid-19 open research dataset." *ArXiv* (2020).

[Xian *et.al*., 2017] Xian, Y., Schiele, B., & Akata, Z. (2017). Zero-shot learning-the good, the bad and the ugly. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4582-4591).

[Yu *et.al*., 2021] Yu, Dongran, Bo Yang, Dayou Liu, Hui Wang, and Shirui Pan. "Recent Advances in Neural-symbolic Systems: A Survey." *arXiv e-prints* (2021): arXiv-2111.