Kashf Journal of Multidisciplinary Research
Vol: 02 - Issue 01 (2025)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

IMPROVING COVID-19 INFORMATION RETRIEVAL BY INTEGRATING SEMANTICS AND CLUSTERING

Khalid Mahmood
Department of Artificial Intelligence, NFC Institute of Engineering and Technology, Multan, Pakistan.

****Muhammad Ahsan Raza***
Department of Information Sciences, University of Education, Lahore, Multan campus 60000, Pakistan.

Ghulam Irtaza
Department of Information Sciences, University of Education, Lahore, 54000, Pakistan.

****Corresponding Author:*** (ahsan.raza@ue.edu.pk)

DOI: (*https://doi.org/10.71146/kjmr236*)

Article Info

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license https://creativecommons.org/licenses/by/4.0

Abstract

A semantic and natural language processing-based information retrieval system can be very supportive of society and experts in fighting the viral disease pandemic, such as COVID-19. Various emerging technologies, including artificial intelligence, machine learning, the semantic web, and big data analytics, are being progressively applied in information retrieval systems to support data retrieval across different disciplines. The healthcare discipline, especially, requires such systems, applications, and technological support to improve the study and quality of research on the chronic viral disease COVID-19. Wisely arranged and precisely retrieved data in healthcare systems is very helpful in providing quality services and making effective decisions to deal with the COVID-19 pandemic. An information retrieval system can be made more effective in producing more precise results by adding semantics such as semantic textual similarity measures, clustering, and the support of ontology. This research presents an improved information retrieval system by integrating semantic clustering and domain ontology. The K-Means is the most used algorithm in various research schemes for clustering terms, but has shortcomings in the context of computing the similarity or close relationship between concepts. The use of semantic similarity measures between data items in the K-Mean clustering procedure provides an effective method for forming more accurate clusters semantically. Further, the integration of the COVID-19 dataset ontology supports computing the relationship between data items more accurately when forming clusters. The semantic clustering and ontology integration can generate more accurate clusters than the clusters formed only through distance calculation between data items. The results of this research proved a higher accuracy level of results retrieved by information retrieval process.

Keywords: *semantic clustering, K-Mean clustering, semantic similarity, COVID-19 ontology, semantic information retrieval.*

I. Introduction

A large amount of newly collected data, including scientific literature, news, and tweets related to the COVID-19 pandemic and its effects on society are generated from various sources [1]. Various natural language processing (NLP) oriented schemes for relevant information retrieval have been introduced and developed for the processing of such textual data to enable the community to not only fight the COVID-19 pandemic but also to deal with the spread, prevention, and cure of such viral diseases [3-4]. Such information retrieval systems are effectively supportive of conversational medical diagnosis systems to find out the affected people by analysing the symptoms, signs, related descriptions, and pre-recorded medical records from the textual data collection [2]. Furthermore, these NLP-based information retrieval or processing systems are also helpful in the grouping of COVID-19 data according to its subcategories, like precautions, prevention, cure, and evolution of the disease [5-8].

The modern development in medical, information, and other technologies is the reason for producing large amounts of data during the COVID-19 pandemic. Similarly, there is a need to process, maintain, integrate, categorise, analyse, and retrieve accurate and relevant information from the data collection. It is also required to organise the data according to the existing and newly introduced complex medical terms, either in hierarchical relationships or proximity of terms. Due to the lack of semantics in already introduced information retrieval systems, there is poor integration and interoperability between information processing systems. In this research work, the integration of semantic clustering and domain ontology is proposed as an information retrieval scheme to represent the information in semantically organised clusters according to the relevance of concepts.

The grouping or clustering of similar data items is the procedure of selecting the most similar data items relevant to a specific concept into a group. The K-Means clustering algorithm is very popular and widely used in various clustering schemes. The selection procedure depends upon the measure of distance (e.g. Euclidean Distance) between data items and the centroid in the K-Mean algorithm [9-10]. Relying only on the distance computation for similarity measure in the K-Mean algorithm is not a good approach because it may not provide effective results due to less attention on the meaning proximity of data items semantically. A semantic similarity approach [11] is added along with the distance measure to compute the meaningful proximity between the data items semantically. Semantic clustering is possible in this way to get more relevant data members in a specified cluster. Along with the addition of a semantic similarity measure in the K-Mean algorithm, the use of the COVID-19 dataset domain ontology for the distance measure between the terms on ontology and the data items in the dataset is proposed in this research work to produce more effective results. To summarise, in this proposed scheme, a COVID-19 dataset ontology is generated and integrated with the semantic K-Mean clustering algorithm, and only textual data is considered to be manipulated hierarchically when using the ontology to measure the semantic similarity.

The rest of the article is organised as the related works are elaborated in the next section. The proposed methodology is described in the next section of related work. After the elaboration of the proposed methodology, the discussion about the model is presented in the second-to-last section, and the work is concluded in the last section, along with the future work.

II. Literature Review

A large variety of work related to NLP-based information retrieval systems has been proposed by various researchers over time, and during the COVID-19 pandemic specifically. These information retrieval schemes work on the sentence or term level similarity measures, while others may incorporate clustering in finding the retrieval results. In this section, various research works are presented related to semantic similarity, clustering, and ontology-based textual information retrieval systems.

A. Clustering Related

The works presented in [12, 16, 17] provide a detailed survey of various approaches, algorithms, and tools used by different researchers. These include agglomerative hierarchical clustering, K-Means, text categorisation, concept mining, information extraction, vector space model for TF-IDF computation, fuzzy clustering, and so on. Furthermore, the literature is studied based on some important factors like similarity measures, the algorithm used for clustering, dataset properties, and the use of ontologies. A semi-supervised K-Means, spherical K-Means, kernel-based K-Means, hierarchical clustering, graph-based document clustering, page rank, and the bisecting K-Means algorithms are adapted in the research work presented by the authors of [18-23], respectively. A rare combination of Multi-level Gaussian Minimum Support and Apriori algorithm is used in graph-based clustering [24]. The drawback of this combination is that sometimes invalid similarity measures are possible in subgraphs, which may directly affect the clustering results. Different performance metrics such as Entropy, Recall, F-measure, Precision, and Silhouette Coefficient are used in different research works for the analysis of results and outcomes [13-15].

In the research work [25], authors have proposed a combination of Adjusted Random Index and Normalised Basic Distance computation as a new evaluation measure for effective clustering. The combination of two clustering algorithms, K-Means and Co-clustering are used in [26] to minimise the space complexity and time. A new clustering algorithm is proposed in [27] for large-scale document collection to optimise the performance in terms of computational time and accuracy in results. The research works [28, 29] presented two clustering algorithms, K-Means and Bisecting K-Means, that are combined with the combination of two similarity measure techniques, inter-similarity and cosine similarity. Authors of [30] used the link structure of Wikipedia to measure the similarity between concepts, but the scheme can be improved with the combination of other techniques. The combination of TRM (Text Relationship Map, GA (Genetic Algorithm and LSI (Latent Semantic Indexing) is used in [31] to provide more semantic similarity measures. In [32], the authors proposed a new scheme for better clustering by combining LSI, VSM, and other information retrieval mechanisms. The scheme proposed in [33] uses macro-averaged F-measure and the K-Mean clustering algorithm, but the module of WSD (Word Sense Disambiguation) is unable to define valid senses.

B. Semantic Similarity Related

The sentence text similarity is introduced in [34, 35] by using an unsupervised learning method. The same technique is proposed in [36, 37], but with a supervised learning methodology to achieve better results. Various research schemes [38, 39, 40, 41] introduced a multi-task learning model based on sentence similarity and embedding techniques to gain high performance over large-scale data.

C. COVID-19 Pandemic Related

Recently, many research works have been produced by researchers on account of the unprecedented coronavirus. In this section, the literature studied specifically about the textual data manipulation related to the COVID-19 pandemic is listed. The proposed method in [42] extracts or finds out the misinformation, rumours, public reaction, and attitudes toward the COVID-19 pandemic from the Twitter dataset. The research work presented in [43] determines the pattern of social media for COVID-19 epidemic data arrangement and categorization to predict the spread of the virus. A textual dataset of public emotional responses recorded and presented in [44], the dataset contains 5k labelled text documents with their corresponding titles. A prediction model was also introduced for the approximation of sentiments or emotions of the dataset. A study and an information processing technique about the COVID-19 policies are presented in [45], which support policymakers, authorities, and society to combat such pandemic disasters. The work presented in [46-49] focuses on the information processing of tweets and messages

on YouTube and Reddit, respectively, related to COVID-19 in various languages. The scheme presented in [50] annotates approximately 1.7k questions from the CORD19 dataset [51], where the label determines the type of question queried within every text document. The questions are categorized into fifteen groups or classes using BERT (Bidirectional Encoder Representations from Transformers) as the base model. Research works presented in [52] proposed a novel methodology for performance improvement in the activities related to question answering systems.

D. Ontology Related

The technological advancements in various disciplines, including medical and computing fields, are introducing new methodologies for the processing and retrieval of information. The use of ontologies is an effective technological advancement for information retrieval systems, especially in the age of infectious pandemics like COVID-19 epidemics [53, 54]. The major role of ontological matching in knowledgebase systems is to identify the semantic relationship between the terms in textual data collection. The ontology knowledge graph makes data integration, the quality of data, and data sharing easy in the system. The domain ontology is the only technology that is used widely for logical reasoning in NLP-based information retrieval systems. The ASMOV (Automated Semantic Matching of Ontologies with Verification) algorithm introduced in [55] is used to iteratively compute the similarity between two concepts based on the feature description of the text. An ontology matching scheme based on the Fuzzy rule in [56] reveals the internal structure of the ontology and the lexical details of data items. The lexical and semantic structural combinations provide effective results in terms of accuracy. An ontology-based information retrieval system [57] was applied in a case study of a clinic system to extract more relevant information. A web application was developed, and an ontology and RDF (Resource Description Framework) schema were created by the use of the Data Interchange Model. In the research work [58], the authors proposed a system in which IDO (Infectious Disease Ontology) interoperates with different medical fields, including biomedical research, clinical care, and public health. The objective of this research is to support the initial discovery of data in future pandemics. The CIDO (Coronavirus Infectious Disease Ontology) supported system was introduced by the authors in [59], which covers a wide variety of domains related to the COVID-19 pandemic, such as diagnosis, a etiology, hindrance, pathogenesis, epidemiology, cure, and treatment. The CIDO ontology provides a logical representation of the COVID-19 details and also provides support for related pandemics. The architecture of CIDO is BFO (Basic Formal Ontology), which covers mostly general classes of ontology. In the CBR (Case-based Reasoning) model [60], NLP is used to construct an ontology for every case separately. The RA (Regression Analysis), along with AI (Artificial Intelligence), is used in [61] for the development of a binary classification model by taking various factors into account, such as the density of the population, wind speed, humidity, and temperature.

The discussed literature shows that domain ontologies have been used in different types of information retrieval systems. It is concluded that a few systems deliver good outcomes on COVID-19 ontologies, which shows their importance in the new pandemic. Along with the use of ontologies, NLP techniques can also be very useful to make the results more effective for any information retrieval system.

III. Proposed Methodology

In this research work, we propose an integration technique that combines both clustering with the semantic similarity measure and domain ontology into a single clustering mechanism for a COVID-19 information retrieval system. In various research works, these techniques are used individually with other NLP techniques, but not used together in any research performed during the COVID-19 pandemic (as discussed in the literature review section). A detailed description of the scheme is given as follows, according to the involved factors. The overall proposed methodology is presented in the following Fig. 1.

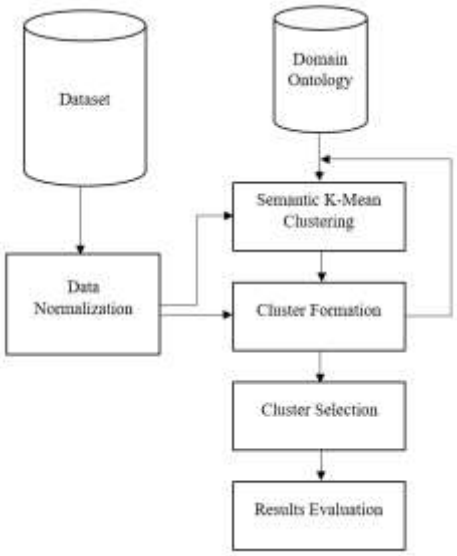

Fig. 1. The proposed methodology

A. Text Preprocessing

The input dataset to the semantic-clustering-based information retrieval system is the collection of text files containing textual data collected during the age of the COVID-19 pandemic. The text preprocessing operations of NLP are performed on the dataset in the first preprocessing phase of experimentation. These operations include tokenization, removal of punctuation from the data, removal of numbers from the data, removal of stop words from the data, POS (Part of Speech) tagging, lemmatization, and stemming. These operations are performed on the dataset for cleaning and transforming unstructured text data to prepare the dataset for further processing in the information retrieval system. The system may produce a new data corpus after performing these operations on the dataset. The resultant dataset contains only meaningful terms or concepts in the corpus.

B. Vector Space Model

The vector space model (VSM) is used for vectorization in this research. It is the most popular and widely used in Information Retrieval systems. In the vector space model, a document is represented as a vector showing some weight value. The value of the weight is computed through the multiplication of Term Frequency (TF) and Inverse Document Frequency (IDF). The calculation of the weights of the term is another significant module in this model. The conversion of text terms into numerical vectors is the major responsibility of this module. There are different mechanisms to transform the textual data into numerical vectors, like Bag of Words, Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec and Doc2Vec. In the proposed VSM scheme, the TF-IDF mechanism is chosen for the calculation of weights.

C. Semantic Clustering

The clustering or grouping of relevant data items or documents has been used in various research schemes, applications and disciplines, including medicine, medical, engineering, biology, as well as data mining or information retrieval. Clustering can also be used to add semantics in the procedure of information retrieval, known as Semantic Information Retrieval, to get more accurate and relevant results of the user query. In this section, the K-Means clustering algorithm is selected for the formation of clusters semantically. The proposed K-Mean algorithm is presented in Fig. 2.

Input: Dataset (X): $X = \{x_1, x_2, x_3, \ldots\ldots x_n\}$ where x_i denotes data item, Number of Clusters (k)

Output: Required Clusters with relevant data items

Initialization of k cluster centroids, $C = \{c_1, c_2, c_3 \ldots\ldots c_k\}$ where each c_k is centroid

Do {

foreach x_i compute the distance

$d_{ij} = d(x_i, c_j)$ for $i = 1, 2, 3 \ldots n$ and $j = 1, 2, 3 \ldots k$

$x_i \in C_j$, $\quad for\ min(d_{ij})$

foreach C_j update new centroid c_j by computing average (mean)

$$c_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

} until (no new centroid found)

Fig. 2. The proposed K-Mean algorithm

The calculation of distance between data points is most cost-effective in this clustering algorithm, which needs special consideration to minimise the overall cost for large textual data. Euclidean Distance calculation is performed to compute the distance between data points and the centroids, as given in Equation (1). A detailed discussion of the semantic similarity measure is given in the coming section.

$$d_{ij} = d(x_i, c_j) \text{ for } i = 1,2,3 \ldots . n \text{ and } j = 1,2,3 \ldots . k \qquad (1)$$

Where x represents the data element and c represents the cluster centroid. This calculation may incur high costs, and it may take more time for large textual data with a large number of features or data items. The process of average distance calculation between data items and the centroids is given in Equation (2) as follows.

$$c_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \qquad (2)$$

The process involving nearest centroid selection and re-computing the average distance will incur a low cost if we fix the number of iterations. The implementation of the K-Means algorithm on various platforms, one by one, is elaborated in the following sections.

D. Semantic Similarity Measure

The computation of semantic similarity to measure the proximity between data items is proposed in this research for the formation of clusters. The proposed methodology is to overcome the drawback of simple distance calculation between data members. The semantic similarity is used to measure the proximity between the concepts with the integration of ontology to get an accurate selection of data members in a cluster. The similarity computation between two terms or concepts depends on the knowledge type normalised for assessment. The taxonomical ontology structure and the term distribution in the data corpus are considered in various schemes. The semantic similarity between terms is computed as the number of links or the length of the path between terms or concepts in the given taxonomy. In this way, the similarity is measured only based on the use of ontology instead of a data corpus as contextual knowledge. In this research, this type of similarity measure between the terms is adapted for the selection of data members of relevant clusters.

A simple method to compute the similarity between two terms t_1 and t_2 Is the shortest length of path calculated between these two terms or the minimum number of connecting links of these two concepts in the nomenclature of ontology, as mathematically represented in the following Equation 3.

$$Sim(t_1, t_2) = min_{edges}(is - a(t_1, t_2)) \qquad (3)$$

Here, the similarity computation is the count of the minimum $IS-A$ links between the two terms. This methodology is improved and represented as another variation. This variation of the similarity measure between two terms was adopted in this research. The similarity is measured based on the number of connecting links between the terms in the taxonomy of ontology. The computation process is mathematically represented in Equation (4).

$$Sim(t_1, t_2) = \frac{2 \times L_3}{L_1 + L_2 + 2 \times L_3} \qquad (4)$$

Here, L_1 and L_2 is the number of $(IS-A)$ links between the two terms $t_1$and t_2 in the taxonomy. The variable L_3 represents the number of $(IS-A)$ links between the real taxonomical predecessor of both the terms and the ontology root as described in [62].

E. Ontology Construction

The conceptual representation of an information retrieval system is made through ontologies. An ontology is the hierarchical representation of concepts in a specific domain, logically related to each other. There are three different ways to obtain knowledge in an information retrieval system: the integration of existing ontology, the addition of new domain-related information to a current ontology, and relying only on the domain-provided information. The construction of a new ontology needs knowledge in the domain, expertise in ontology engineering, and skills to operate the software used for ontology construction. The token identification, the representation of synonyms, the concept identification, the understanding of hierarchical relationships among concepts, and the acquisition of rules are some major subtasks needed to learn for the integration of ontology in an information retrieval system.

In this research, a COVID-19 ontology is constructed for the integration with the clustering procedure used in the proposed information retrieval system. A sample COVID-19 pandemic hierarchy is represented in the following Fig. 3, for the proposed ontological taxonomy.

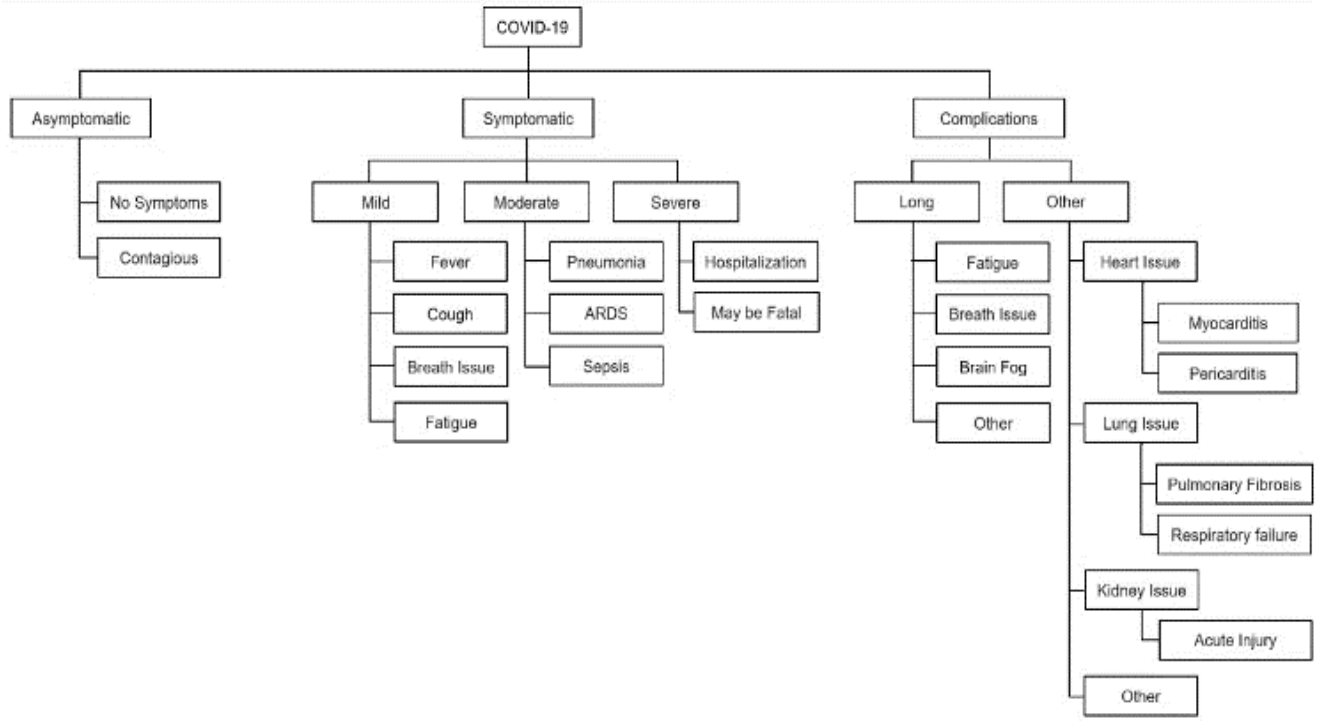

Fig. 3. A sample COVID-19 hierarchy of proposed ontological taxonomy

The proposed COVID-19 ontology is constructed based on pandemic types and their complications. In

Fig. 4, there is a representation of concept hierarchy implemented in the Protégé ontology construction tool. It is recommended that the substantiation of matching classes, the knowledge comprehensiveness, the consistency of concepts, and the knowledge base extension without altering the semantics of the current information retrieval system be mandatory before the construction of the ontology. The structure of the ontology is formed in a classified operational arrangement. The proposed COVID-19 ontology is constructed in a top-down fashion, in which the concept belonging to the most general class is placed at the top level of the hierarchy, and the concepts belonging to its subclasses are arranged at successive levels of the hierarchy. Various concepts belong to a class, and major concepts among those concepts represent classes.

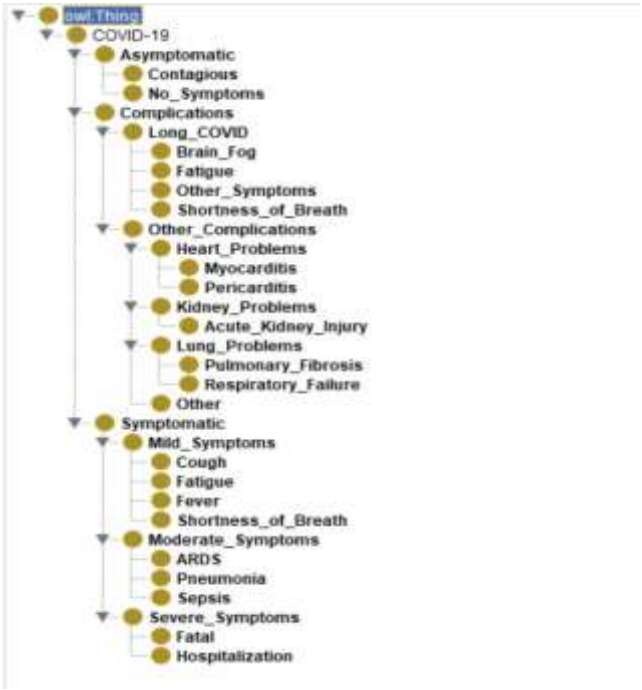

Fig. 4. The concept hierarchy of the proposed COVID-19 ontology

The term COVID-19 is a primary concept in this ontological structure. The related concepts of the pandemic are further classified into subclasses and their concepts in the relevant hierarchy. All the concepts at auxiliary levels all are derived from the primary concept, and by following this concept-driving mechanism, the whole concept hierarchical structure of the ontology is constructed. The Protégé 5.6.3 version of the ontology construction tool is used to develop the domain knowledge base. Fig. 5 shows the visualization of $IS-A$ hierarchical structure of symptomatic classification of the COVID-19 pandemic was developed in Protégé as an example. The Protégé is a powerful tool that provides support for ontology management, also supports the classes of OWL, and has a wide set of knowledge modelling assemblies, which makes the ontology construction easy.

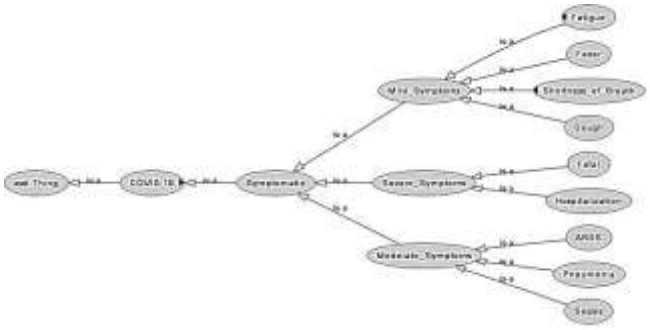

Fig. 5. The Visualisation of $(is - a)$ link hierarchical structure of symptomatic concepts

In the ontology construction phase of this research, the relationship between concepts and the affected humans is the properties of these concepts. These properties are used to define the relationship between data items and ontological concepts. These properties have $IS - A$ relationship between the terms in the dataset and the concepts. In Fig. 6, an example COVID-19 pandemic-infected case is represented in an ontological concept hierarchy.

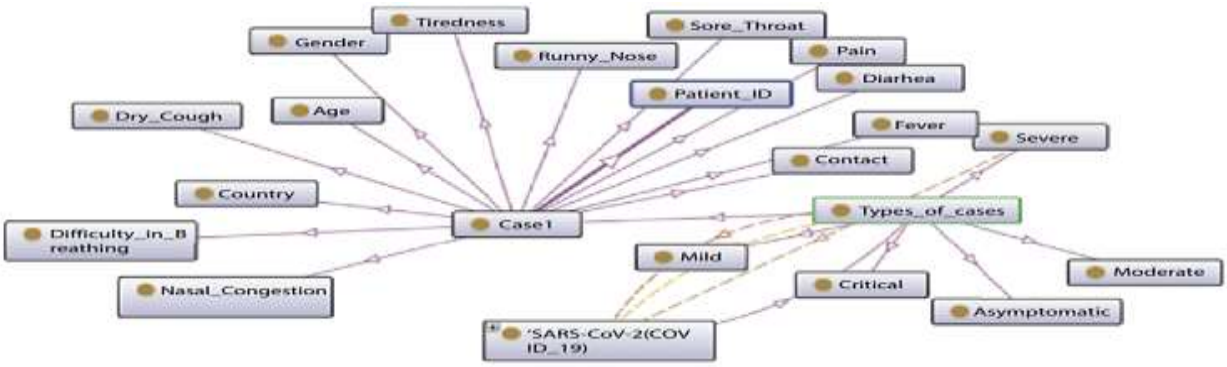

Fig. 6. Ontological representation of the concept hierarchy of an example

IV. Discussion

All these major phases, including preprocessing, vector space model, semantic similarity measures, semantic clustering, and ontology construction (as discussed in the previous section), are involved in the proposed semantic clustering and ontology integration of the COVID-19 information retrieval system. A block diagram of semantic clustering with the utilisation of domain ontology is presented in the following Fig. 7, which shows a detailed description of the semantic clustering procedure.

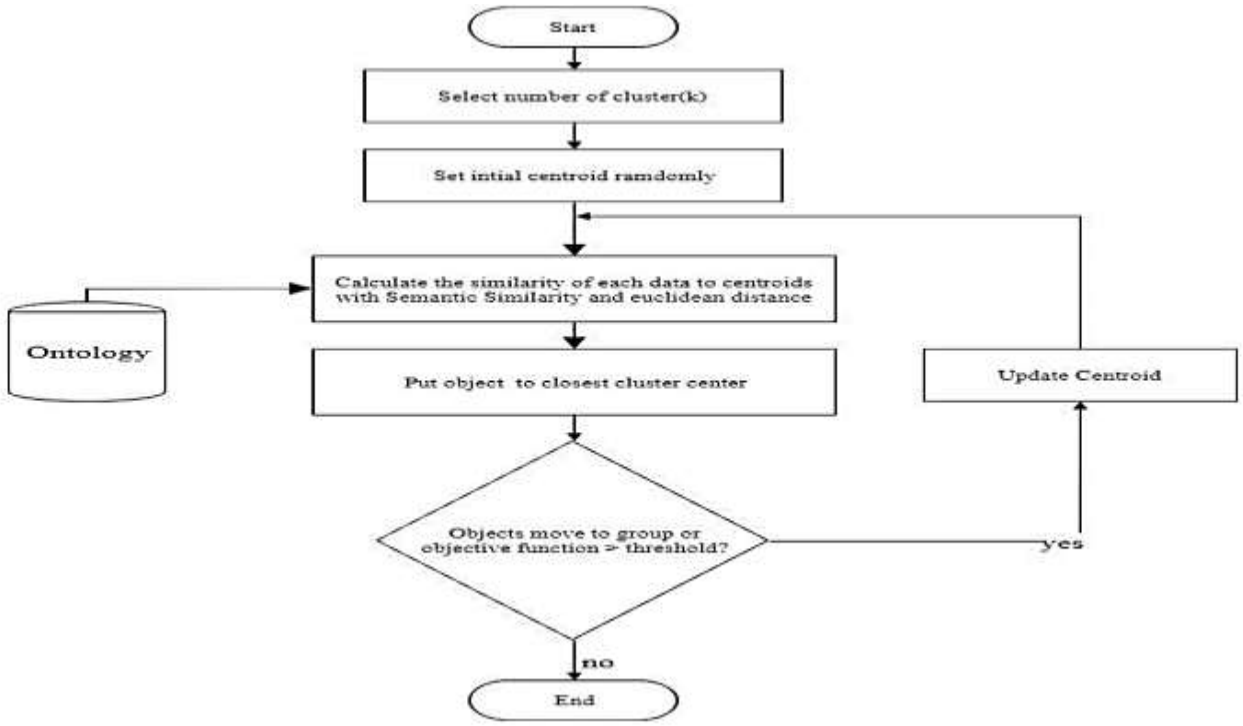

Fig. 7. Semantic clustering using domain ontology and semantic similarity

In the proposed model, the K-Means clustering algorithm is selected to perform the clustering process as discussed in the previous section. In the K-Mean algorithm, first, the number of clusters is determined to define initial centroid values for each cluster. Initially, centroid values are selected randomly. After the selection of the initial centroid value, the similarity distance is computed between each of the data items and the centroid.

The data items are categorized according to their high similarity with a centroid in a specific cluster; the data items are assigned to the cluster as its data member accordingly. The process of the similarity computation is proposed to be performed using the Euclidean distance measure and the semantic similarity computation (as presented in the semantic similarity measure section). After the allocation of cluster members, the next phase is the convergence of cluster outcomes, in which the results of newly formed clusters are compared with the results of clusters formed in the previous iteration. If the results are the same, then the results are converged, otherwise not converged. The new centroid value is computed for each cluster, and the next iteration is performed. The procedure is repeated until there is no change in the membership of clusters.

The experimentation can be performed to measure the accuracy and relevance of the results of semantic clustering based on semantic similarity with ontology integration. The results can be computed in terms of Accuracy, Precision, Recall, F-score, and other metrics. The proposed model with integration of semantics and clustering obviously produces more effective results. The authors planned to implement an enhanced and semantic COVID-19 information retrieval system based on this strategy. The comparison with other relevant approaches and the detailed evaluation are planned to be presented in future.

V. Conclusion and Future Work

The COVID-19 pandemic is a severe viral disease which affected the globe in previous years. Various research works presented and played their significant role during the pandemic in various dimensions to fight against the disease. An effective information retrieval and management system has its unique significance to play its role in decision-making to fight such situations. An enhanced and effective semantic clustering with domain ontology integration for the COVID-19 information retrieval system is proposed in this research article. An ontology construction in the domain of COVID-19 is proposed, and semantic clustering based on the semantic similarity of concepts with the integration of the proposed ontology is presented in this article. A textual dataset related to the COVID-19 pandemic is best suited for the application of this proposed model. The K-Mean clustering algorithm is adapted to perform the central process of cluster formation of concepts or text. The domain ontology-based and semantic similarity-clustering are integrated to develop an intelligent COVID-19 information retrieval system. In future, the developed ontology and semantic clustering technique will be tested to obtain optimal results using different NLP-based information retrieval systems.

References:-

[1] Guo, X., Mirzaalian, H., Sabir, E., Jaiswal, A., & Abd-Almageed, W. (2020). CORD19STS: COVID-19 semantic textual similarity dataset (arXiv:2007.02461) [Dataset]. arXiv. https://arxiv.org/abs/2007.02461

[2] Selvam, M., & Amane, N. (2025, March). Enhancing question answering systems with semantic networks and NLP techniques. In 2025 International Conference on Computing for Sustainability and Intelligent Future (COMP-SIF) (pp. 1–6). IEEE.

[3] Ben Abacha, A., & Zweigenbaum, P. (2015). MEANS: A medical question-answering system combining NLP techniques and semantic web technologies. Information Processing & Management, 51(5), 570–594.

[4] Zendaoui, F., Hidouci, W. K., & Rouhani, S. (2022). Uncertainty identification in microblogs. Journal of Optimization in Industrial Engineering, 15(1), 301–309.

[5] [5] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36.

[6] Syed-Abdul, S., Fernandez-Luque, L., Jian, W. S., Li, Y. C., Crain, S., Hsu, M. H., Wang, Y. C., Khandregzen, D., Chuluunbaatar, E., Nguyen, P. A., et al. (2013). Misleading health-related information promoted through video-based social media: Anorexia on YouTube. Journal of Medical Internet Research, 15(2), e30.

[7] Sakai, H., & Lam, S. S. (2025). Large language models for healthcare text classification: A systematic review. arXiv preprint arXiv:2503.01159.

[8] Mohammed, S. M., Madaan, V., Resen, R. D., Sharma, N., & Hassen, O. A. (2025). Text categorization for information retrieval using NLP models. Journal of Cybersecurity & Information Management, 15(2).

[9] Gupta, M. K., & Chandra, P. (2022). Effects of similarity/distance metrics on k-means algorithm with respect to its applications in IoT and multimedia: A review. Multimedia Tools and Applications, 81(26), 37007–37032.

[10] Prasad, R. K., Chakraborty, S., & Sarmah, R. (2023). Impact of distance measures on partition-based clustering method—An empirical investigation. International Journal of Information Technology, 15(2), 627–642.

[11] Harispe, S., Ranwez, S., & Montmain, J. (2022). Semantic similarity from natural language and ontology analysis. Springer Nature.

[12] Jacksi, K., Dimililer, N., & Zeebaree, S. R. M. (2015). A survey of exploratory search systems based on LOD resources. In Proceedings of the 5th International Conference on Computing & Informatics (pp. 501–509). Sintok, Malaysia.

[13] Jacksi, K., Dimililer, N., & Zeebaree, S. R. M. (2016). State of the art exploration systems for linked data: A review. International Journal of Advanced Computer Science and Applications, 7(11), 155–164. https://dx.doi.org/10.14569/IJACSA.2016.071120

[14] Patil, H., & Thakur, R. (2018). A semantic approach for text document clustering using frequent itemsets and WordNet. International Journal of Engineering & Technology, 7, 102. https://doi.org/10.14419/ijet.v7i2.9.10220

[15] Ibrahim, R., Zeebaree, S., & Jacksi, K. (2019). Survey on semantic similarity based on document clustering. Advances in Science, Technology and Engineering Systems Journal, 4(5), 115–122. https://doi.org/10.25046/aj040515

[16] Jacksi, K., & Badiozamany, S. (2015). General method for data indexing using clustering methods. International Journal of Science and Engineering, 6(3), 641–644.

[17] Zafar, A., Awais, M., & Aftab, M. A. (2018). Ontology-based document data analysis.

[18] Raza, Asif, Salahuddin, & Inzamam Shahzad. (2024). Residual Learning Model-Based Classification of COVID-19 Using Chest Radiographs. Spectrum of Engineering Sciences, 2(3), 367–396.

[19] Khan, Zia, Saif Ur Rehman Khan, Omair Bilal, Asif Raza, and Ghazanfar Ali. "Optimizing Cervical Lesion Detection Using Deep Learning with Particle Swarm Optimization." In 2025 6th International Conference on Advancements in Computational Sciences (ICACS), pp. 1-7.

[20] Avanija, J., & Ramar, K. (2015). Semantic similarity-based clustering of web documents using fuzzy C-means. International Journal of Computational Intelligence and Applications, 14. https://doi.org/10.1142/S1469026815500157

[21] Adel, A., Al-Zebari, S. R. M., Jacksi, K., & Selamat, A. (2019). ELMS–DPU ontology visualization with Protégé VOWL and Web VOWL. Journal of Advanced Research in Dynamical and Control Systems, 11(1), 478–485.

[22] Afreen, S., & Srinivasu, D. B. (2017). Semantic-based document clustering using lexical chains.

[23] Al-Khasawneh, Mahmoud Ahmad, Asif Raza, Saif Ur Rehman Khan, and Zia Khan. "Stock Market Trend Prediction Using Deep Learning Approach." Computational Economics (2024): 1-32.

[24] Wei, T., Lu, Y., Chang, H., Zhou, Q., & Bao, X. (2015). A semantic approach for text clustering using WordNet and lexical chains. Expert Systems with Applications, 42(4), 2264–2275. https://doi.org/10.1016/j.eswa.2014.10.023

[25] Blokh, I., & Alexandrov, V. (2017). News clustering based on similarity analysis. Procedia Computer Science, 122, 715–719. https://doi.org/10.1016/j.procs.2017.11.428

[26] Manzali, Y., Barry, K. A., Flouchi, R., Balouki, Y., & Elfar, M. (2025). An innovative clustering approach utilizing frequent item sets. Multimedia Tools and Applications, 84(10), 7835–7861.

[27] Elsayed, A., Mokhtar, H., & Ismael, O. (2015). Ontology-based document clustering using MapReduce. International Journal of Database Management Systems, 7. https://doi.org/10.5121/ijdms.2015.7201

[28] Raza, Asif , Soomro, M. H., Shahzad, I., & Batool, S. (2024). Abstractive Text Summarization for Urdu Language. Journal of Computing & Biomedical Informatics, 7(02).

[29] Conrad, J. G., & Bender, M. (2016). Semi-supervised events clustering in news retrieval (pp. 21–26).

[30] Khan, Z., Hossain, M. Z., Mayumu, N., Yasmin, F., & Aziz, Y. (2024, November). Boosting the prediction of brain tumor using two stage BiGait architecture. In 2024 International Conference on Digital Image Computing: Techniques and Applications (DICTA) (pp. 411-418). IEEE.

[31] Kolhe, S. R., & Sawarkar, S. D. (2017). A concept-driven document clustering using WordNet. In 2017 International Conference on Nascent Technologies in Engineering (ICNTE) (pp. 1–5). IEEE. https://doi.org/10.1109/ICNTE.2017.7947888

[32] Jacksi, K. (2019). Design and implementation of E-Campus ontology with a hybrid software engineering methodology. Scientific Journal of University of Zakho, 7(3), 95–100.

[33] S. U. R. Khan, A. Raza, I. Shahzad and G. Ali, "Enhancing Concrete and Pavement Crack Prediction through Hierarchical Feature Integration with VGG16 and Triple Classifier Ensemble," 2024 Horizons of Information Technology and Engineering (HITE), Lahore, Pakistan, 2024, pp. 1-6.

[34] Mousavi, N., Scerri, S., & Auer, S. (2017). Semantic similarity-based clustering of license excerpts for improved end-user interpretation. https://doi.org/10.1145/3132218.3132224

[35] Kavitha, C., Sadhasivam, S., & Kiruthika, S. (2014). Semantic similarity-based web document classification using artificial bee colony (ABC) algorithm. WSEAS Transactions on Computers, 13, 476–484.

[36] Khan, U. S., Ishfaque, M., Khan, S. U. R., Xu, F., Chen, L., & Lei, Y. (2024). Comparative analysis of twelve transfer learning models for the prediction and crack detection in concrete dams, based on borehole images. Frontiers of Structural and Civil Engineering, 18(10), 1507-1523.

[37] Ilievski, F., Shenoy, K., Chalupsky, H., Klein, N., & Szekely, P. (2024). A study of concept similarity in Wikidata. Semantic Web, 15(3), 877–896.

[38] Sumathy, M. K. L., & Chidambaram, D. (2016). A hybrid approach for measuring semantic similarity between documents and its application in mining the knowledge repositories. International Journal of Advanced Computer Science and Applications, 7(8). https://doi.org/10.14569/IJACSA.2016.070831

[39] Khan, M. A., Khan, S. U. R., Haider, S. Z. Q., Khan, S. A., & Bilal, O. (2024). Evolving knowledge representation learning with the dynamic asymmetric embedding model. Evolving Systems, 15(6), 2323-2338.

[40] Bafna, P., Pramod, D., & Vaidya, A. (2016). Document clustering: TF-IDF approach. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (pp. 61–66). https://doi.org/10.1109/ICEEOT.2016.7754750

[41] Rezaie, V., & Parnianifard, A. (2022). A new intelligent system for diagnosing tumors with MR images using type-2 fuzzy neural network (T2FNN). Multimedia Tools and Applications, 81(2), 2333–2363.

[42] Shahzad, I., Khan, S. U. R., Waseem, A., Abideen, Z. U., & Liu, J. (2024). Enhancing ASD classification through hybrid attention-based learning of facial features. Signal, Image and Video Processing, 18(Suppl 1), 475-488.

[43] Mondal, S., Gurushanker, A., Loganath, M., Chowdhury, R., Karthik, S., Kalinathan, L., & Palani, S. (2025). Exploring methodologies for computing sentence similarity in natural language. Advances in Data and Information Sciences: Proceedings of ICDIS 2024 (Vol. 2, pp. 251–1193).

[44] Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., St John, R., Constant, N., Tar, C. (2018). Universal sentence encoder. arXiv preprint arXiv:1803.11175.

[45] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364.

[46] Khan, U. S., & Khan, S. U. R. (2025). Boost diagnostic performance in retinal disease classification utilizing deep ensemble classifiers based on OCT. Multimedia Tools and Applications, 84(19), 21227-21247.

[47] Khan, S. R., Asif Raza, Inzamam Shahzad, & Hafiz Muhammad Ijaz. (2024). Deep transfer CNNs models performance evaluation using unbalanced histopathological breast cancer dataset. Lahore Garrison University Research Journal of Computer Science and Information Technology, 8(1).

[48] Zhou, Y., Chen, C., & Huang, G. (2024, September). Sentence similarity model based on semantics and syntax. In 2024 5th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE) (pp. 261–264). IEEE.

[49] Liu, X., He, P., Chen, W., & Gao, J. (2019). Improving multi-task deep neural networks via knowledge distillation for natural language understanding. arXiv preprint arXiv:1904.09482.

[50] Subramanian, S., Trischler, A., Bengio, Y., & Pal, C. J. (2018). Learning general purpose distributed sentence representations via large-scale multi-task learning. arXiv preprint arXiv:1804.00079.

[51] Asif, Sohaib, Yi Wenhui, Saif ur-Rehman, Qurrat ul-ain, Kamran Amjad, Yi Yueyang, Si Jinhai, and Muhammad Awais. "Advancements and prospects of machine learning in medical diagnostics: unveiling the future of diagnostic precision." Archives of Computational Methods in Engineering 32, no. 2 (2025): 853-883.

[52] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

[53] Amur, Z. H., Kwang Hooi, Y., Bhanbhro, H., Dahri, K., & Soomro, G. M. (2023). Short-text semantic similarity (STSS): Techniques, challenges and future perspectives. Applied Sciences, 13(6), 3911.

[54] O. Bilal, Asif Raza, S. ur R. Khan, and Ghazanfar Ali, “A Contemporary Secure Microservices Discovery Architecture with Service Tags for Smart City Infrastructures ”, VFAST trans. softw. eng., vol. 12, no. 1, pp. 79–92, Mar. 2024

[55] Levy, S., & Wang, W. Y. (2020). Cross-lingual transfer learning for COVID-19 outbreak alignment.

[56] Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set. JMIR Public Health and Surveillance, 6(2), e19273.

[57] Lin, H., & Bu, N. (2022). A CNN-based framework for predicting public emotion and multi-level behaviors based on network public opinion. Frontiers in Psychology, 13, 909439.

[58] S. ur R. Khan, Asif. Raza, Muhammad Tanveer Meeran, and U. Bilhaj, “Enhancing Breast Cancer Detection through Thermal Imaging and Customized 2D CNN Classifiers”, VFAST trans. softw. eng., vol. 11, no. 4, pp. 80–92, Dec. 2023. DOI: 10.21015/vtse.v11i4.1684

[59] Alsudias, L., & Rayson, P. (2020). COVID-19 and Arabic Twitter: How can Arab world governments and public health organizations learn from social media?

[60] Kruspe, A., Haeberle, M., & Zhu, X. X. (2020). Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic.

[61] HUSSAIN, S., Raza, A., MEERAN, M. T., IJAZ, H. M., & JAMALI, S. (2020). Domain Ontology Based Similarity and Analysis in Higher Education. IEEEP New Horizons Journal, 102(1), 11-16.

[62] Aggarwal, J., Rabinovich, E., & Stevenson, S. (2020). Exploration of gender differences in COVID-19 discourse on Reddit.

[63] Serrano, J. C. M., Papakyriakopoulos, O., & Hegelich, S. (2020). NLP-based feature extraction for the detection of COVID-19 misinformation videos on YouTube.

[64] Mahmood, F., Abbas, K., Raza, A., Khan,M.A., & Khan, P.W. (2019). Three Dimensional Agricultural Land Modeling using Unmanned Aerial System (UAS). International Journal of Advanced Computer Science and Applications (IJACSA) [p-ISSN : 2158-107X, e-ISSN : 2156-5570], 10(1).

[65] GitHub, Inc. (2020). Open-source data: COVID-Q [Dataset]. GitHub. https://github.com/JerryWei03/COVID-Q

[66] Friel, R., Belyi, M., & Sanyal, A. (2024). Ragbench: Explainable benchmark for retrieval-augmented generation systems. arXiv preprint arXiv:2407.11005.

[67] GitHub, Inc. (2020). Open-source data: COVID-QA [Dataset]. GitHub. https://github.com/deepset-ai/COVID-QA

[68] Wang, N., Issa, R. R., & Anumba, C. J. (2022). NLP-based query-answering system for information extraction from building information models. Journal of Computing in Civil Engineering, 36(3), 04022004.

[69] Raza, A., & Meeran, M. T. (2019). Routine of Encryption in Cognitive Radio Network. Mehran University Research Journal of Engineering and Technology [p-ISSN: 0254-7821, e-ISSN: 2413-7219], 38(3), 609-618.

[70] Kaur, P., & Khamparia, A. (2014). Review on medical care ontologies. International Journal of Scientific Research, 3(12), 677–680.

[71] Bain, D., & Dutta, B. (2024). Systematic analysis of COVID-19 ontologies. In Research Conference on Metadata and Semantics Research (pp. 74–91). Springer.

[72] Jean-Mary, Y. R., Shironoshita, E. P., & Kabuka, M. R. (2009). Ontology matching with semantic verification. Journal of Web Semantics, 7(3), 235–251.

[73] Fernández, R. S., Velasco, J., Marsa-Maestre, I., & Lopez-Carmona, M. (2012). Fuzzy Align: A fuzzy method for ontology alignment. In Proceedings of the International Conference on Knowledge Engineering and Ontology Development (KEOD 2012) (pp. 98–107).

[74] Ibrahim, A. M., Hashi, H. A., & Mohamed, A. A. (2013). Ontology driven information retrieval for healthcare information system: A case study. International Journal of Network Security & Applications, 5(1), 61–69.

[75] Babcock, S., Cowell, L. G., Beverley, J., & Smith, B. (2020). The Infectious Disease Ontology in the age of COVID-19. OSF Preprints.

[76] He, Y., Yu, H., Ong, E., et al. (2020). CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. Scientific Data, 7, 181. https://doi.org/10.1038/s41597-020-0523-6

[77] Meeran, M. T., Raza, A., & Din, M. (2018). Advancement in GSM Network to Access Cloud Services. Pakistan Journal of Engineering, Technology & Science [ISSN: 2224-2333], 7(1).

[78] Oyelade, O. N., & Ezugwu, A. E. (2020). COVID-19: A natural language processing and ontology-oriented temporal case-based framework for early detection and diagnosis of novel coronavirus. Preprints, 2020050171.

[79] Pirouz, B., Shaffiee Haghshenas, S., Shaffiee Haghshenas, S., & Piro, P. (2020). Investigating a serious challenge in the sustainable development process: Analysis of confirmed cases of COVID-19 through a binary classification using artificial intelligence and regression analysis. Sustainability, 12(6), 2427.

[80] Sánchez, D., Batet, M., Isern, D., & Valls, A. (2012). Ontology-based semantic similarity: A new feature-based approach. Expert Syst