 doi:10.20944/preprints202005.0171.v2

A Case-Based Reasoning Framework for Early Detection and Diagnosis of Novel Coronavirus

Olaide N. Oyelade[1], Absalom E. Ezugwu[2]

[1]Department of Computer Science, Ahmadu Bello University Zaria, Nigeria

[2]School of Computer Science, University of KwaZulu-Natal, King Edward Avenue, Pietermaritzburg Campus, Pietermaritzburg, 3201, KwaZulu-Natal, South Africa

Corresponding authors emails: {olaide_oyelade@yahoo.com, ezugwua@ukzn.ac.za}

Abstract:

Coronavirus, also known as COVID-19, has been declared a pandemic by the World Health Organization (WHO). At the time of conducting this study, it had recorded over 1.6 million cases while more than 105,000 have died due to it, with these figures rising on a daily basis across the globe. The burden of this highly contagious respiratory disease is that it presents itself in both symptomatic and asymptomatic patterns in those already infected, thereby leading to an exponential rise in the number of contractions of the disease and fatalities. It is therefore crucial to expedite the process of early detection and diagnosis of the disease across the world. The case-based reasoning (CBR) model is an effective paradigm that allows for the utilization of cases' specific knowledge previously experienced, concrete problem situations or specific patient cases for solving new cases. This study therefore aims to leverage the very rich database of cases of COVID-19 to solve new cases. The approach adopted in this study employs the use of an improved CBR model for state-of-the-art reasoning task in classification of suspected cases of Covid19. The CBR model leverages on a novel feature selection and semantic-based mathematical model proposed in this study for case similarity computation. An initial population of the archive was achieved with 68 cases obtained from the Italian Society of Medical and Interventional Radiology (SIRM) repository. Results obtained revealed that the proposed approach in this study successfully classified suspected cases into their categories at an accuracy of 97.10%. The study found that the proposed model can support physicians to easily diagnose suspected cases of Covid19 base on their medical records without subjecting the specimen to laboratory test. As a result, there will be a global minimization of contagion

 doi:10.20944/preprints202005.0171.v2

rate occasioned by slow testing and as well reduce false positive rates of diagnosed cases as observed in some parts of the globe.

Keywords: COVID-19**,** coronavirus, case-based reasoning, ontology, natural language processing.

1.0 Introduction

The novel coronavirus disease, also referred to as COVID-19, was first identified in China in December 2019. The virus has so far affected 213 countries and territories around the world and 2 international conveyances and is now considered a major global health concern due to its pathogenicity and widespread distribution around the world. The Covid-19 virus is a highly contagious respiratory disease that has raced around the globe since it first emerged in China in late December 2019. According to the World Health Organisation's official reports on COVID-19 [13], by April 21, 2020, it had affected more than 2,482,598 million people and caused more than 170,484 deaths, with a total of 652,543 recovered. Considering the exponential growth in the confirmed and death cases of COVID-19, this has expedited efforts by the scientific and research community in proposing and developing several novel epidemiological model approaches to mitigate the spread of the COVID-19 outbreak.

A number of mathematical and statistical models have been developed recently to critically analyze the transmission pattern of the ongoing COVID-19 and other related disease outbreaks [14, 15, 16, 17, 18, 19]. It is equally important to recognize all the different epidemiological contributions towards estimating the transmission dynamics of the virus, but most of the existing proposed models are parameter dependent and they rely mainly on multiple assumptions [20] for them to be effective. Moreover, because during an outbreak of any epidemic, it is often not easy and reliable to estimate parameters using real data sets, which are not readily available for experimental testing of such proposed models [21, 22]. Furthermore, in most of the reported model parameter settings, one can discover that rather than using the actual parameter values that seem close enough to the real-world values derived from the statistical properties of the actual data sets, the authors of those models opted to use hypothesized parameter values. However, the use of hypothesized parameters in this case is highly limited because it does not fit the data very well [20]. Therefore, considering the aforementioned challenges associated with the current existing mathematical and statistical epidemiological models, it would be very difficult to attribute any high predictive accuracy level for using these

 doi:10.20944/preprints202005.0171.v2

models to correctly estimate and forecast the exponential growth of COVID-19 outbreaks. As it stands for now, despite all these measures and attractive modeling proposals, the virus has maintained its capacity to spread exponentially from country to country and continent to continent stretching the functionality and capability of even the most robust healthcare systems of so many countries.

Although many related artificial intelligence (AI) based proposed studies in the literature appear to be well-designed for the tasks of handling the current coronavirus pandemic in terms of estimating confirmed cases and forecasting the speed of COVID-19 spread, these models may deteriorate in performance and accuracy due to their heavy reliance on many inaccurate decision variables and imprecise parameter estimations. Thus, it is assumed that the aforementioned limitations can lead to conflicting forecasting outcomes, which may invariably lead to unsatisfactory and imprecise results. This would obviously have a negative impact on public health planning and policy making. To overcome the limitations of the aforementioned existing epidemiological and AI-based model approaches, the current paper presents a promising alternative diagnostic and forecasting framework with the aim to achieve more accurate results and avoid the previous limitations by combining the strengths of ontology-based natural language processing with case-based reasoning for early detection and diagnosis of the novel coronavirus pandemic. The rich database of cases of confirmed COVID-19 supports the adoption of case-based reasoning (CBR) paradigm as an authentic reasoning structure for improving diagnosis.

Case based reasoning (CBR) is an artificial intelligence paradigm that has proven to be effective in medical systems, and also exploits the similarity of cases in its knowledge base in providing a solution to a new case or problem. Case(s) retrieval that is closely related with the new case is usually computed using different similarity computational models like Euclidean distance which have been adopted by different researches. However, CBR systems all have the challenge of features extraction and formalization. Furthermore, the choice of selecting the best distance measure model for computing similarity of cases is a problem demanding optimal solution considering the sensitivity of medical cases. Case-based reasoning means using old experiences to understand and solve new problems. In case-based reasoning, a reasoner remembers a previous situation similar to the current one and uses that to solve the new problem [59]. Case-based reasoning (CBR) and expert systems have a long tradition in artificial intelligence. CBR has been formulated since the late 1970s. CBR is an approach for problem solving and learning of humans and computers [60]. Case-based reasoning is useful in problem solving and automation of learning by an agent. Because

 doi:10.20944/preprints202005.0171.v2

empirical evidence has shown that reasoning with CBR is more powerful, this has made reasoning by re-using past cases a powerful and frequently applied way to solve problems for humans. A very important feature of case-based reasoning is its coupling to learning, and its strong association with machine learning [61]. Ben-Bassat, et al. [62] enumerated some features of CBR, and these include: cases that present similar symptoms and findings results from same faults/disease, and "Nearest Neighbor" algorithm is used to identify unknown diagnosis from the known. CBR is beneficial when compared to RBR. CBR avoids the knowledge based acquisition bottleneck of RBR, it compiles past solutions, mimics the diagnostic experience of human experts, avoids past mistakes, interprets rules, supplements weak domain models, facilitates explanation, supports knowledge acquisition and learning, and exploits the database of solved problems so as to learn.

In this paper, we introduce the concept of hybridising natural language processing, ontology learning and artificial intelligence techniques to the most important challenges in responding to the novel coronavirus pandemic. Consequently, the main goal of this study is to apply the concept of natural language processing (NLP) for ontology learning and population task before using an improved CBR technique to the problem of classifying cases of COVID-19 as either positive or negative even when the disease is still in its early stage of manifestation in the presented case. An NLP model for feature extraction of presented case was designed and implemented. The originality of the current study lies in the robustness and efficiency of the sentence-level extraction of feature-value pair for all a-priori declared features. Furthermore, the case retrieval similarity metric applied to the proposed NPL-based CBR framework contributes to the interesting performance of the proposed system. Specifically, the technical contributions of this study are as follows:

i. Design of an ontology learning algorithm for feature extraction and mapping from suspected cases of Covid19.
ii. Proposal of a novel mathematical model for semantic-based and feature based case similarity computation.
iii. Incorporation of the proposed mathematical model into an improved CBR framework.
iv. Implementation of CBR framework which allows for the detection or classification of suspected cases of COVID-19 as either positive or negative.

 doi:10.20944/preprints202005.0171.v2

The remainder of the paper is organized into six sections, namely: the related works, proposed approach, experimentation, results, discussion, and conclusion. The related works section presents a comprehensive review of related studies on COVID-19. In Section 3, a detail of the approach proposed for the CBR framework is presented, while Section 4 discusses the experimentation and system configuration for the experimentation. In Section 5, we present a comparison of the performance of the proposed approach with some related studies, and then conclude the study in Section 6.

2.0 Related Works

This section consists of two parts; the first part covers the detailed review of all the current related work that has so far been implemented to handle the ongoing novel coronavirus.

2.1. Related Work

In recent times, artificial intelligence (AI) has been considered as a potentially powerful tool in the fight against many evolving pandemics such Ebola hemorrhagic fever (2014-2016), Swine flu (2002-2003), SARS (2002-2003), Middle East respiratory syndrome coronavirus (MERS-CoV) (2012-present), and novel coronavirus (COVID-19) (2019-ongoing). Regarding the ongoing 2019-2020 novel coronavirus pandemic, dozens of research efforts have emerged and most of the published papers focused on the important of harnessing artificial intelligence technologies to curb the global COVID-19 Pandemic. This section provides a selective review of recent articles that have discussed the many significant contributions of the application of AI technologies in the fight against COVID-19, as well as the current constraints on these contributions. Specifically, in [1], six areas where artificial intelligence technologies have emerged as key solutions to combatting coronavirus were identified. These areas include: i) early warnings and alerts, ii) tracking and prediction, iii) data dashboards, iv) diagnosis and prognosis, v) treatments and cures, and vi) social control. Therefore, most of the subsequent discussions presented in this section are focused on investigating to what extent AI has been partly or fully utilized in combatting the spread of the aforementioned pandemic. The selected review discussions presented in this section only cover those articles that have been published in a peer reviewed journal. Preprinted articles are outside the scope of the current review discussion.

In [2] the analysis of confirmed cases of COVID-19 through a binary classification using artificial intelligence and regression analysis was investigated. In their study, the authors employed the binary classification modelling with

 doi:10.20944/preprints202005.0171.v2

group method of data handling type of neural network as one of the artificial intelligence methods of accurately predicting confirmed cases of the COVID-19 epidemic. The study chose the Hubei province in China for their model construction. For the input and output variables, some important factors such as maximum, minimum, and average daily temperature, city density, relative humidity, and wind speed, were considered as the input dataset, while the number of confirmed cases was selected as the output dataset for 30 days. Moreover, the outcome of the investigation revealed that the proposed binary classification model was able to provide a higher performance capacity in predicting the confirmed cases in the province. In addition, the analysis of the results also showed that certain weather conditions based on the input variables, namely relative humidity with an average of 77.9% had a positive impact on the confirmed cases and maximum daily temperature with an average of 15.4 °C had a negative impact on the confirmed cases.

Mohammed, et al. [3] presented the application of two optimization metaheuristic techniques to enhance the predictive performance accuracy of the proposed adaptive neuro-fuzzy inference system that is used for estimating and forecasting the number of confirmed cases of novel coronavirus in the upcoming ten days based on previously confirmed cases that were recorded in China. The developed hybrid metaheuristic based adaptive neuro-fuzzy inference system comprised of Adaptive Neuro-Fuzzy Inference engine and two metaheuristic algorithms namely, flower pollination algorithm and salp swarm algorithm. The enhanced flower pollination algorithm was utilized by the author to train the neuro-fuzzy inference system by optimizing its parameters, while the salp swarm algorithm was incorporated as a local search method to enhance the quality of the solution obtained by the model. The results of the model implementation show that it has a high capability of predicting the number of confirmed cases within the projected ten days. It was further established that the hybrid system, when compared with other methods, obtained more superior performance accuracy in in terms of the following performance metrics: root mean square error, mean absolute error, mean absolute percentage error, root mean squared relative error, and coefficient of determination.

Ting, et al. [4] explored the potential application of four inter-related digital technologies combating the wide spread of the novel coronavirus. These technologies include the Internet of Things, big-data analytics, Artificial Intelligence and blockchain. The authors in their work [4] presented some valid reasons why the four aforementioned digital technologies can be employed to augment the already strained traditional based public-health strategies for tackling COVID-19. Some of the traditional based public healthcare strategies that have been put in place and are constantly

 doi:10.20944/preprints202005.0171.v2

being used across the globe include: (1) monitoring, surveillance, detection and prevention of COVID-19; and (2) mitigation of the impact to healthcare indirectly related to COVID-19. The authors further suggested that digital technologies can be helpful in the following ways. The Internet of Things technology can be used to provide a platform that allows public-health agencies access to data for monitoring the COVID-19 pandemic. The big data technology can be very useful in providing opportunities for performing modelling studies of viral activity and for guiding an individual country's healthcare policymakers to enhance preparation for the outbreak. The blockchain technology can be vital in the manufacturing and distribution of COVID-19 vaccines once they are available. Similarly, blockchain can be utilized to facilitate the distribution of patients' regular medication to the local pharmacy or patients' doorstep. The AI and deep learning technology can be used to enhance the detection and diagnosis of COVID-19. Further, the utilization of various AI-based triage systems could potentially alleviate the clinical load of physicians.

Vaishya, et al. [5] in their study highlighted the significant roles that some of the new technologies such as artificial intelligence, Internet of Things, Big Data and Machine Learning are likely to play in the fight against the new diseases and also the possible forecasting of any pandemics. The authors in [5] focused on presenting a brief review regarding the utilization of artificial intelligences platforms as a decisive technology to analyze, prepare us for prevention and fight against COVID-19 and any other similar pandemics. In their findings, seven significant application areas of artificial intelligence technology were identified for tackling the spread of COVID-19 disease. These areas as mentioned in [5] include: early detection and diagnosis of the infection, monitoring the treatment, projection of cases and mortality, development of drugs and vaccines, reducing the workload of healthcare workers, and prevention of the disease. Furthermore, the technology was also identified as having the capability to detect clusters of cases and predict the possible location of the virus spread through collecting and analyzing all previous data.

Leung and Leung [6] presented a discussion on the way forward in terms of crowdsourcing data to mitigate epidemics. The authors surveyed different and varied sources of possible line lists for COVID-19. The sources considered by the authors include data clearing houses or secondary repositories and official websites or social media accounts of various Health Commissions at the provincial and municipal levels in mainland China. Some of the main bottlenecks attributed to the process of crowdsourcing were linked to the rigorous tasks involved in carefully collating as much relevant data as possible, sifting through and verifying the data, extracting intelligence to forecast and inform outbreak strategies, and thereafter repeating this process in iterative cycles to monitor and evaluate progress [6].

 doi:10.20944/preprints202005.0171.v2

However, a possible methodological breakthrough in alleviating these challenges would be to develop and validate algorithms for automated bots to search through cyberspaces of all sorts, by text mining and natural language processing to expedite these processes. Next, we present a brief discussion of some applications of CBR to healthcare with a specific focus on its utilization for analysis, prediction, diagnosis, and recommending treatment for patients.

The CBR is an appropriate methodology to apply in the diagnosis and treatment of wide range of health issues. Research in CBR has grown to an extent, starting from the early exploration in the medical field by Koton [7], Bareiss [8] in the late 1980s and Gierl, et al. [9] in the late 1990s. However, there are still some associated shortcomings with the design and implementation of CBR, especially in the adaptation mechanism. Blanco, et al. [10] reported the results of a systematic review of CBR application to the health sector. In their work, the authors proposed some enhancement procedures that could be applied to overcome some of the limitations of CBR, which is focused on preparing the data to create association rules that help to reduce the number of cases and facilitate learning of adaptation rules.

CBR has equally received noticeable attention in the aspect of disease predictions and diagnosis. In [11], a hybrid implementation of neural networks and case-based reasoning was proposed for the prediction of chronic renal disease among the Colombian population. The neural network-based classifier which was trained with the demographic data and medical care information of two population groups was developed to predict whether a person is at risk of developing chronic kidney disease. The result of the classifier showed that about 3,494,516 people were identified as being at risk of developing chronic renal disease in Colombia, which in this case is 7% of the total population.

Benamina, et al. [12] proposed the integration of fuzzy logic and data mining technique to improve the response time and the accuracy of the retrieval step of case-based reasoning of similar cases. The Fuzzy CBR proposed in [12] is composed of two complementary parts, namely, the part of classification by fuzzy decision tree realized by Fispro and the part of case-based reasoning realized by the platform JColibri. The main function of fuzzy logic is to reduce the complexity of calculating the degree of similarity that can exist between diabetic patients who require different monitoring plans. The authors compared their results with some existing classification methods using accuracy as performance metrics. The experimental result that was generated by the proposed system revealed that the fuzzy decision tree is very effective in improving the accuracy for diabetes classification and hence improving the retrieval step of CBR reasoning.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

Table 1 presents a concise summary of all published work on mathematical modeling, statistical modeling and simulation based literature on COVID-19 that appeared in Web of Science database. Each paper information is represented based on the article authors' details in column two, paper title in column three and reference in column four. Overall, sixteen publications were retrieved from the Web of Science database.

Table 1: Summary of all published related modeling and simulation based work on COVID-19 from Web of Science

SN	Author	Title	ref
1	Rao, Arni S. R. Srinivasa; Krantz, Steven G.; Kurien, Thomas; et al.	Model-based retrospective estimates for COVID-19 or coronavirus in India: continued efforts required to contain the virus spread	[23]
2	Buonomo, Bruno	Effects of information-dependent vaccination behavior on coronavirus outbreak: insights from a SIRI model	[24]
3	Kim, Soyoung; Kim, Yae Jean; Peck, Kyong Ran; et al.	School Opening Delay Effect on Transmission Dynamics of Coronavirus Disease 2019 in Korea: Based on Mathematical Modeling and Simulation Study	[25]
4	Hellewell, Joel; Abbott, Sam; Gimma, Amy; et al.	Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts	[26]
5	Jia, Jiwei; Ding, Jian; Liu, Siyu; et al.	Modeling the control of COVID-19: Impact of policy interventions and meteorological factors	[27]
6	Choi, Sunhwa; Ki, Moran	Estimating the reproductive number and the outbreak size of COVID-19 in Korea	[28]
7	Huang, Rui; Liu, Miao; Ding, Yongmei	Spatial-temporal distribution of COVID-19 in China and its prediction: A data-driven modeling analysis	[29]
8	Gostic, Katelyn; Gomez, Ana C. R.; Mummah, Riley O.; et al.	Estimated effectiveness of symptom and risk screening to prevent the spread of COVID-19	[30]
9	Wang, Chuanyi; Cheng, Zhe; Yue, Xiao-Guang; et al.	Risk Management of COVID-19 by Universities in China	[31]
10	Roosa, Kimberlyn; Lee, Yiseul; Luo, Ruiyan; et al.	Short-term Forecasts of the COVID-19 Epidemic in Guangdong and Zhejiang, China: February 13-23, 2020	[32]
11	Yang, Shu; Cao, Peihua; Du, Peipei; et al.	Early estimation of the case fatality rate of COVID-19 in mainland China: a data-driven analysis	[33]
12	Jiang, Xiangao; Coffee, Megan; Bari, Anasse; et al.	Towards an Artificial Intelligence Framework for Data-Driven Prediction of Coronavirus Clinical Severity	[34]
13	Yin, Fulian; Lv, Jiahui; Zhang, Xiaojian; et al.	COVID-19 information propagation dynamics in the Chinese Sina-microblog	[35]
14	Zhou, Weike; Wang, Aili; Xia, Fan; et al.	Effects of media reporting on mitigating spread of COVID-19 in the early phase of the outbreak	[36]
15	Yang, Chayu; Wang, Jin	A mathematical model for the novel coronavirus epidemic in Wuhan, China	[37]
16	Rong, Xinmiao; Yang, Liu; Chu, Huidi; et al.	Effect of delay in diagnosis on transmission of COVID-19	[38]

 doi:10.20944/preprints202005.0171.v2

17	Hou, Can, Jiaxin Chen, Yaqing Zhou, Lei Hua, Jinxia Yuan, Shu He, Yi Guo et al.	The effectiveness of the quarantine of Wuhan city against the Corona Virus Disease 2019 (COVID-19): well-mixed SEIR model analysis	[39]
18	Prem, Kiesha, Yang Liu, Timothy W. Russell, Adam J. Kucharski, Rosalind M. Eggo, Nicholas Davies, Stefan Flasche et al.	The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study	[40]
19	Ho, Dean	Addressing COVID-19 Drug Development with Artificial Intelligence.	[41]
20	Shi, Feng, Jun Wang, Jun Shi, Ziyan Wu, Qian Wang, Zhenyu Tang, Kelei He, Yinghuan Shi, and Dinggang Shen	Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for Covid-19	[42]
21	Kim, Donghyun, Soyoung Hong, Sungwoo Choi, and Taeseon Yoon.	Analysis of transmission route of MERS coronavirus using decision tree and Apriori algorithm	[43]

3.0 Proposed Approach

A detail presentation of the methods adopted and adapted in this study is covered in this section: an overview of the entire approach, feature extraction using a natural language processing technique, formalism of cases in the proposed case-based reasoning (CBR) method, and lastly the CBR engine.

3.1 An Overview of the Approach

The proposed NLP-Ontology-CBR method accepts a text-based patient file as input for processing of status of the case. Figure 1 presents an illustration of all procedures applied to the inputs denoted by a case file. The case file is passed as input into an NLP Text-2-Features module. This module leverages on some natural language processing operations to synthetically and semantically extract tokens from the case file. The extracted case features are further passed to the domain-based feature extraction component which maps each extract feature at the previous layer to domain-based features. The extracted and mapped features are formalized using description logic (DL) based on a knowledge representation format to allow for efficient computational operations in the CBR engine. Finally, the formalized features are passed on to the CBR-engine as a new case (nc) that support the application of the reasoning paradigm of CBR.

The pipeline of information flow and processing described in Figure 1 was therefore adapted to detect the case of positive COVID-19 patient from early stage to the advance stage. A further discussion in the following subsections details the components of the framework.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

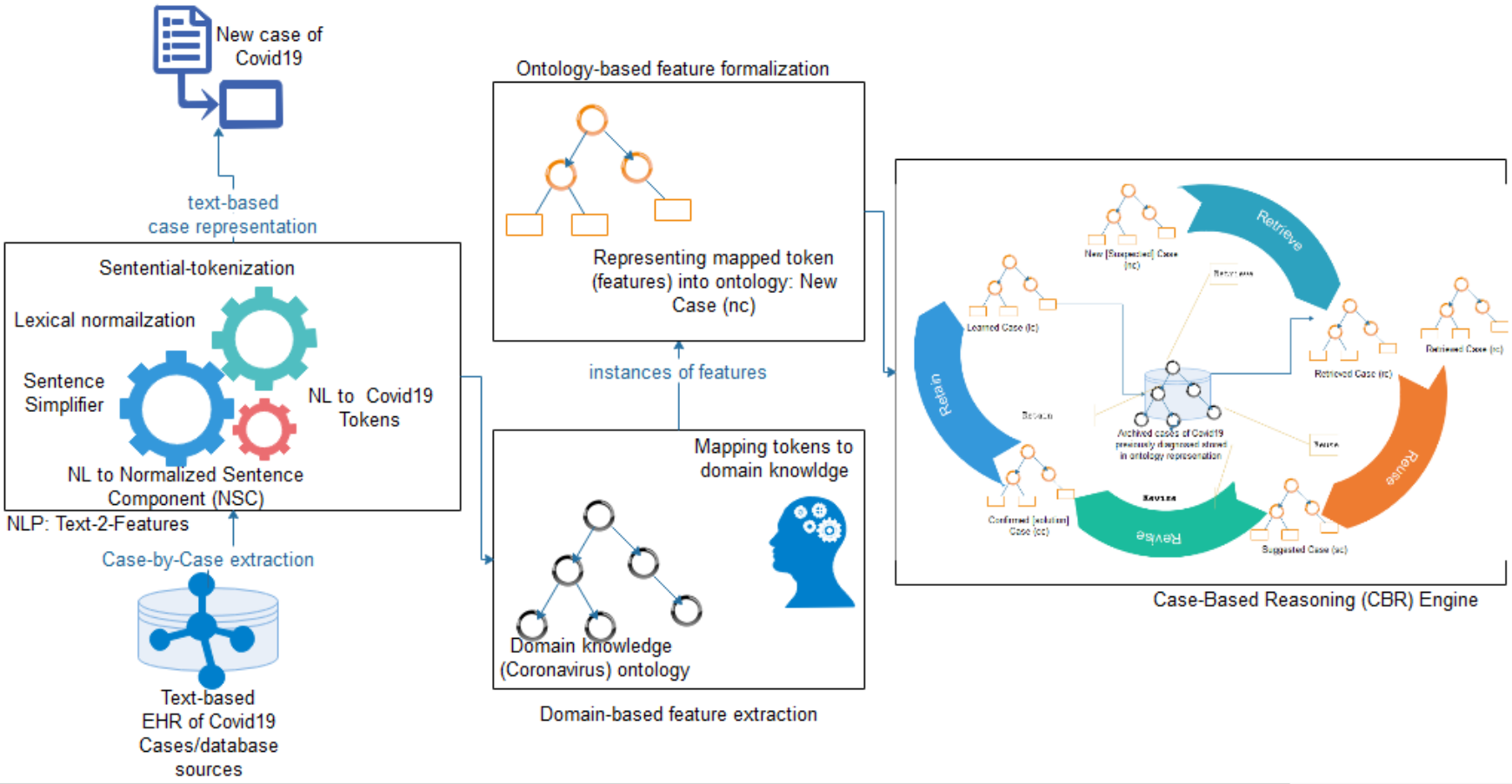

Figure 1: An overview of the proposed framework using case-based reasoning (CBR) model to classify new cases of coronavirus (COVID-19) as either positive or negative

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

3.2 The NLP Method for Feature Extraction

The field of natural language processing technique is a very interesting and relevant aspect of artificial intelligence (AI) with a wide range of applications to medicine and even the large number of text-based documents on the internet. Moreover, electronic health record (EHR) systems are now pervasive and are provided as services to other automated healthcare delivery systems. The NLP method for feature extraction described in this section adopts some components and algorithms from Dasgupta, et. al. [44]. The ontology learning method is widely used for mining information from natural language text and generating an ontology representation of the mined data. Such ontology representation is to provide formal expressivity and a platform for reasoning with such NLP-text document. Although this study assumes a similar procedure, we implemented a skeletal of the entire procedure.

Figure 2 shows the modified model of a patient text-based medical record natural language processing (NLP) and features extraction pipeline. The model is called a pipeline because of its approach of processing raw file-based text (in English language) through different procedures which eventually yields the feature (Covid-Fs) for further processing in the CBR-engine.

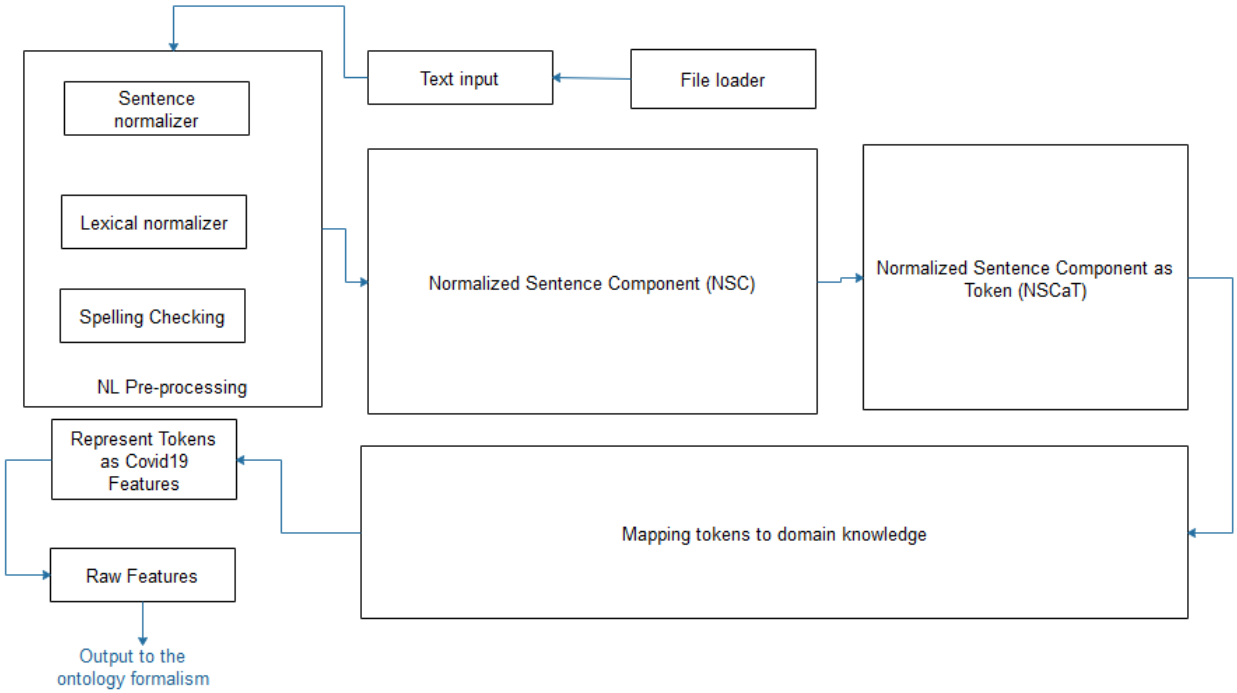

Figure 2: A patient text-based medical record natural language processing (NLP) and feature extraction architectural pipeline

The following is a breakdown of the components of the NLP processing pipeline as shown in Figure 2:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

a. File loader and Text input (FLTI)
b. NL Pre-processing: Spelling checking, Lexical normalizer, Sentence normalizer
c. Normalized Sentence Component (NSC)
d. Normalized Sentence Component as Token (NSCaT)
e. Mapping tokens to domain knowledge (MTDK)
f. Represent Tokens as COVID-19 Features (RTCF)
g. Raw Features Buffer (RFB)

File Loader and Text Input (FLTI): The FLTI is a very simple component with support for file format and safety authentication, file loading and text-content unloaded into a buffer.

NL Pre-processing (NL-P): The second component consists of other sub-modules named Spelling Checking, Lexical normalizer, and Sentence normalizer which does a pre-processing of the buffered text in FLTI layer. Generally, the NL-P is aimed at carrying out operations like spell-corrector, tokenization, sentence boundary detector, text singularizer, POS-tagger, co-reference resolver, and named-entity recognizer (NER) by leveraging on Stanford coreNLP toolkit [45]. Our approach of applying NL-P to the buffered text in FLTI was to allow the spell-corrector to scan through the complete buffer and correct wrongly spelt words and furthermore to allow for efficient and intelligent mining of features from the buffered text - the improved output of FLTI. This was then converted into a token of sentential forms (SF) in list format and then sorted according to their appearance in the original document. In each SFs, we attempted to normalize each plural form of its constituents into a singular form through the use of a singularizer. These SFs were extracted from buffered text using sentence boundary detector and annotated with POS-tagging, and the SFs were preserved in an orderly manner to sustain the semantics of health records. Meanwhile, due to the translation task of the raw text to ontology format, we further employed NER models to identify and mark entities and thereafter their instances which form the elements of taxonomy-box (TBox) and assertion box (Abox) respectively in the resulting ontology. Once the SFs had been pre-processed, we applied them to the next sub-module named lexical normalizer (LN). The use of LN in our study is simply for identification of quantifiers and special symbols (like >, <, =, +, -, and other medical related symbols which may hold meaning in the usage) of subject/objects appearing in the SFs. Our approach in LN allows for such quantifiers/numeric representations and symbols to be normalized into normal forms supportive of the token-to-feature translation in RTCF component of Figure 2. The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

role of applying the sentence normalizer (SN) is to ensure that very difficult sentences are broken down to simple forms so that an element of SFs, say sf_i, is normalized into simpler forms assuming the template of the NSC component to be discussed later. Hence, the resulting simplified sentences of sf_i replace it in SFs.

Normalized Sentence Component (NSC): Based on the structural formation of a sentence in English language, a particular template or syntax was described by Dasgupta, et. al. [44] in their study. We adopted two of the templates, namely the simple and complex sentences as listed in the following:

$$\underline{Q_1}\ \underline{M_1^*}\ S\ \text{is-a}\ \underline{Q_2}\ \underline{M_2^*}\ O$$

$$\underline{Q_1}\ \underline{M_1^*}\ S\ Cl_1\ \text{IS-A}\ \underline{Q_2}\ \underline{M_2^*}\ O_1\ Cl_2\ \text{IS-A}\ \underline{Q_3}\ \underline{M_3^*}\ O_2$$

Q Under-lined notation indicates optional component with at most 1 occurrence in the template e.g. quantification

M* Under-lined notation with asterisk (*) indicates 0 or more consecutive occurrences in the template e.g. adjectives.

Q_1 Subject quantifier that includes lexical variations of the set: a, an, the, some, all.

Q_2 Object quantifier that includes lexical variations of the set: the, some, all.

Q_3 Object quantifier that includes lexical variations of the set: the, some, all.

M Subject/object/verb modifier; value is restricted to the set: Noun, Adjective, Adverb, Numerical, and Gerund

S Subject; value is restricted to the set: fNN, NNP25, JJ, RB, VBGg

O Object; value is restricted to the set: fNN, NNP, JJ, RB, VBGg

IS-A Denotes all possible lexical variations.

Cl_1 and **Cl_2** signifies IS-A clausal token and all its variations ‘which’, ‘who’, ‘whose’, ‘whom’, ‘that’.

Finally, we ensured that all the sentences in SFs were adapted to the template described above and then we applied their Template-Fitting algorithm to all elements of SFs.

Normalized Sentence Component as Token (NSCaT): The CBR-engine to be described in Sub-section 3.4 does not expect input in sentential format but tokenized features which maintains its sentence form syntax and semantics. Therefore, each sf_i in SFs are further tokenized into a list of raw (un-normalized features) tokens in the form of t_{ij} such

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

that ***i*** represents the position of the sentence in SFs and ***j*** represents the position of the token in the sf_i of SFs that is being processed. The output of NSCaT is therefore an irregular 2D array of raw tokens.

Mapping Tokens to Domain Knowledge (MTDK): We assumed that not all the tokens from NSCaT are correctly represented based on the domain knowledge. As a result, we proposed a MTDK layer which was aimed at mapping each token in the NSCaT to its correct recognized name in the domain. We relied strongly on Wordnet (WordNet) and the domain-based lexicon model in this study as shown in Figure 10. The role of the Wordnet lexicon is to generate all likely synonyms of each t_{ij} in NSCat. This therefore means that each t_{ij} indexes into a sub-array of its synonyms. Thereafter, our mapping algorithm aligns each t_{ij} to its respective sub-array.

Represent Tokens as COVID-19 Features (RTCF): The output of MTDK is further refined to assume the standard feature categorization and typing as listed in Table 2. The implication of this is that we attempted to extract known features of COVID-19 from the output of MTDK and assigned their values as illustrated in Figure 3.

Raw Features Buffer (RFB): This last component simply buffers the output of raw features collected from previous layers. The RFs buffered in RFB are then translated into ontology formalism described in Section 4.2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

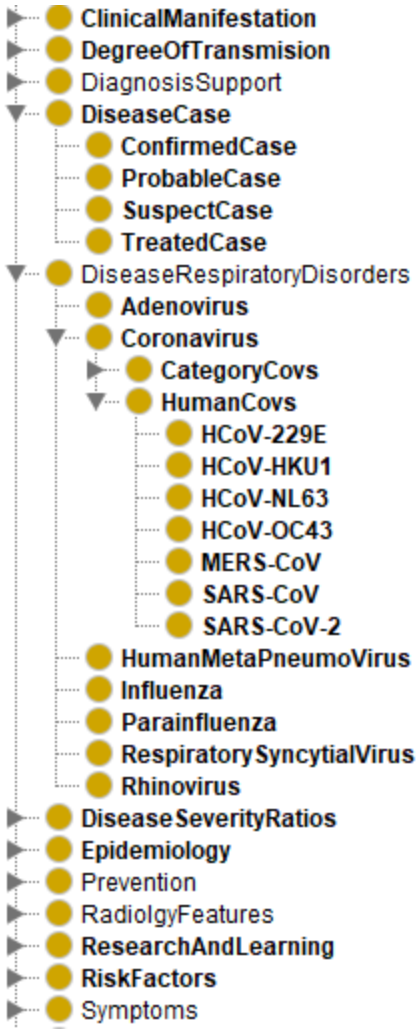

Figure 3: A lexicon of terminologies representing domain knowledge of COVID-19 in addition to symptoms, treatment, epidemiology, disease case status, and other relevant concepts in the domain

The features described in Table 2 were based on recent studies on COVID-19 that were discussed by Michelen, et al. [46] and Yang, et al. [33].

Table 2: A summary of categorization of coronavirus clinical-based features to be extracted by the domain-based feature extractor

Feature category	Feature Name	Description of feature	Feature calibration
Epidemiological	Sex	Gender of patient	Male/Female
	Basic Reproduction	-	range: 1.5–3.5
	Mortality rate	-	3%
	Incubation time	-	4.8 ± 2.6 days
	Age of the deaths	Median age of death was 75	Range: 48 and 89

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

	BMI	Body mass index	23.75 (4.54)
	Height	-	167 (11.75)
	Weight (kg)	-	65.92 (18.75)
	Age	Patient current age	45.11 ± 13.35
Symptom	Cough	Observed in less than half of the mild cases in the largest included study and in two thirds of cases.	Y | N
	Fever	The most frequent symptom for mild and moderate cases	< 39.1 °C
	Anosmia	Stronger predictor of COVID-19 than self-reported fever amongst people in the community	Y | N
	Pneumonia	Found in severe cases	Y | N
	Acute respiratory distress syndrome (ARDS)	Found in severe cases. Different forms of ARDS are distinguished based on the degree of hypoxia. When PaO2 is not available, a ratio SpO2/FiO2 ≤ 315 is suggestive of ARDS	Y | N
	Organ failure	Found in severe cases	Y | N
	Dyspnea	Rare	Y | N
	Nausea and vomiting	Rare	Y | N
	Headache	More frequent in severe cases	Y | N
	Diarrhoea	-	Y | N
	Respiratory tract infections	-	Y | N
	Shortness of breath	-	Y | N
	Snotty	-	Y | N
	Rhinorrhea	-	Y | N
	Gastrointestinal symptoms	-	Y | N
	Muscle pain	-	Y | N
	Loss of appetite	-	Y | N
	PaO_2	kpa, range80-100	Numeric value
	SaO	Ranges between ≥95%	Numeric value
	Loss of smell	Strong prediction	Y | N
	Heart rate	Beats per minute	Around 88.63
	Systolic pressure	Measured in mmgH	Around 129.98
	Diastolic pressure	Measured in mmgH	Around 81.69
	Fatigue	-	Y | N
	Expectoration	Most common	Y | N
	Septic Shock	Deemed the most critical of them all	Y | N
	Sepsis Shock	Deemed the most critical of them all	Y | N
	Sore throat	Pain in any pair of breast	Y | N
	pH	Hydrogen ion concentration	Around 7.11
	Temperature (°C)	-	>= 37.86
	Pharyngeal pain	-	Y | N
	Chest pain/tightness	Not frequent, with less than 5% of mild cases	Y | N
	Abdominal pain	-	Y | N
Exposure/Travel History (Spatial/Location)	Contact with people	-	Y | N
	Stay in areas with community spread	-	Y | N
Comorbidity (diseases)	Cardio-cerebrovascular	-	Y | N
	Digestive system	-	Y | N

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

	Endocrine diseases	-	Y | N
	Runny nose	-	Y | N
	Malignant tumor	-	Y | N
	Neural system	-	Y | N
	Respiratory system diseases	-	Y | N
Laboratory Tests	neutrophil (× 10^9per L)	-	range 1.8–6.3
	Leucocyte (× 10^9per L)	-	range 3.5–9.5
	Lymphocyte (× 10^9 per L)	-	range 1.1–3.2
	platelet (× 10^9 per L)	-	range 125–350
	Blood coagulation	-	
	Active partial thrombin time	-	range 22–36
	Prothrombin time	-	range 10–13.5
	D-dimer	-	range <0.55
	albumin	-	range 35–57
	ALT (IU/L)	-	range 0–64
	AST ((IU/L)	-	range 8–40
	Total bilirubin (μmol/L)	-	range 4.7–24
	Urea nitrogen (mmol/L)	-	range2.6–7.5
	Creatinine (μmol/L)	-	range 41–73
	CK (mmol/L)	-	range 40–200
	LDH (U/L)	-	range12–250
	Serum Lactate (mmol/L)	-	range >2
	Glucose (mmol/L)	-	range3.9–6.1
	Coagulopathy	-	Y | N
	C-reactive protein (mg/L)	Infection-associated	range 0.0–6.0
	Procalcitonin	Elevation to evidence of COVID-19	-
Treatment	Oxygen therapy	-	Y | N
	Antifungal treatment	-	Y | N
	Antiviral treatment	-	Y | N
	Extracorporeal membrane oxygenator (EMO)	-	Y | N
	Glucocorticoids	-	Y | N
	Antibiotic treatment	-	Y | N
	Intensive care unit (ICU)	-	Y | N
	Noninvasive ventilation (NIV)	-	Y | N
	Invasive mechanical ventilation (IMV)	-	Y | N
Radiological	Pulmonary infiltration	-	Y | N
	Air bronchogram	-	Y | N

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

	Centrilobular nodules	-	Y | N
	Tree-in-bud	-	Y | N
	Reticular pattern	-	Y | N
	Subpleural linear opacity	-	Y | N
	Bronchial dilatation	-	Y | N
	Cystic change	-	Y | N
	Lymphadenopathy	-	Y | N
	Pleural effusion	-	Y | N

3.3 Ontology-based Formalization of Extracted Features

In this stage of our proposed CBR-framework, we processed the raw features buffered in the RBF component of Figure 2 into ontology formalism. Recall that the proposed framework relies on the CBR paradigm to reasoning over the cases presented to it. Hence, each case was modeled using a formalism supporting computational reasoning operation. Figure 4 demonstrates an illustration of a case denoted by Case N. We assumed that based on clinical protocols of COVID-19, a case representation must have a relationship to Diagnosis Case (Suspected, Confirmed, Presumed status); Symptoms (as listed in Table 1); Epidemiology (as listed in Table 1); Radiology/Laboratory manifestations (as listed in Table 1); Clinical Diagnosis (Mild, Acute, Severe); and Treatment (as listed in Table 1). Each case of COVID-19 extracted by the NLP pipeline described in Figure 2 was formalized into this structure as illustrated in Figure 4. The Diagnosis Case entity assumes a 1-1 relationship with every case; Symptoms however presents with a one-many 1 to many (1-M) relationship for each case; also, the Epidemiology entity allows each case to manifest one-many (1-M) relationship; the Radiology/Laboratory manifestations entity also presents each case in a one-many (1-M) relationship given the number of lab tests and radiological operations that might be exercised for each case; Clinical Diagnosis, however, allows for one-one (1-1) relationship due to the fact that a case can only assume one of the states listed in clinical diagnoses; and finally, the Treatment entity allows for one-many (1-M) because one case may respond to one or more treatment/therapy administered to it.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

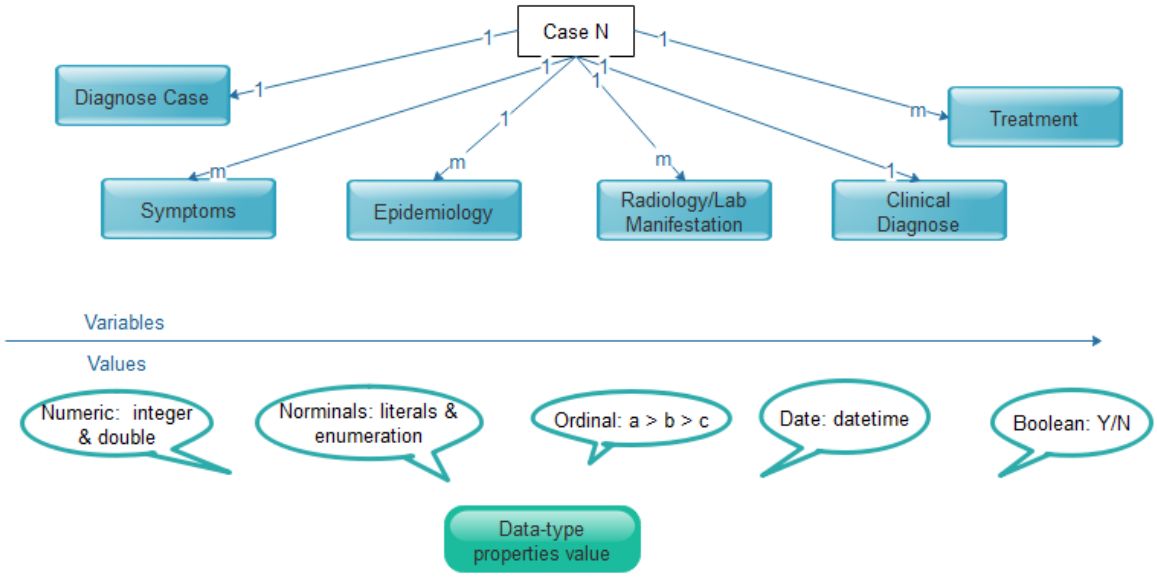

Figure 4: A formal representation tokens (features) of a new case (nc) of coronavirus (COVID-19)

Moreover, each entity illustrated in Figure 4 consists of variables/features which are expected to have values. For instance, considering the Symptom entity, it may have variables/features like ***Cough, Fever, Chest Pain*** and others. Each of those variables are expected to take values from a particular data type. Hence we further illustrated the class of data typing each of the entity may draw denotes the values of its variables. Potential data types as captured in Figure 4 are numeric, nominal, ordinals, datetime, and Boolean (which forms the largest representation for most values of variables in the representation).

3.4 The CBR model

All previous stages of the proposed CBR-based framework may be classified as data/input pre-processing and formalization operations. However, the main reasoning task is embodied in the CBR engine to be described in this section. Meanwhile, we shall first present a brief description of some status or clinical types of COVID-19 based on clinical presentation [47]:

Mild case: Upper respiratory symptoms such as pharyngeal congestion, sore throat, and fever for a short duration or asymptomatic infection; Positive RT-PCR test for SARS-CoV-2; no abnormal radiographic and septic presentation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

Moderate case: Mild pneumonia; symptoms such as fever, cough, fatigue, headache, and myalgia; and absence of complications and manifestations related to severe conditions.

Severe case: A case presenting with mild or moderate clinical features described above; rapid breath (≥70 breaths per min for infants aged <1 year; ≥50 breaths per min for children aged >1 year); hypoxia; lack of consciousness, depression, coma, convulsions; dehydration, difficulty feeding, gastrointestinal dysfunction; myocardial injury; elevated liver enzymes; coagulation dysfunction, rhabdomyolysis, and any other manifestations suggesting injuries to vital organs.

Critical illness case: Respiratory failure with need for mechanical ventilation, persistent hypoxia that cannot be alleviated by inhalation through nasal catheters or masks; septic shock; organ failure that needs monitoring in the ICU, and acute respiratory distress syndrome (ARDS). Cases presenting with ARDS may show:

i. Mild ARDS: 200 mmHg < PaO2/FiO2 ≤ 300 mmHg. In not-ventilated patients or in those managed through non-invasive ventilation (NIV) by using positive end-expiratory pressure (PEEP) or a continuous positive airway pressure (CPAP) ≥ 5 cmH2O.
ii. Moderate ARDS: 100 mmHg < PaO2/FiO2 ≤ 200 mmHg.
iii. Severe ARDS: PaO2/FiO2 ≤ 100 mmHg.

These clinical types of COVID-19 have been described to allow for their use in the CBR engine which will be described below.

The CBR method is a reasoning paradigm that depends on a knowledge base of archived cases that have been proven and tested with valid solutions for handling new cases/problems which may share similar features with those archived. As earlier stated, this study builds on this paradigm to carry out the detection and diagnoses of COVID-19 in patients manifesting symptoms of the disease and those presenting with asymptomatic cases. Figure 12 illustrates our concept of the CBR engine embedded in Figure 8. The major components of the model are similar to the conventional CBR model which usually consists of the RETRIEVE, REUSE, REVISE, and RETAIN steps (4Rs). In addition, the model shows the knowledge base of archived cases which allow for carrying out computational reasoning on the new case presented. The distinctiveness of our proposed CBR model lies in its ability to model its cases using ontology

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

formalism and as well to measure similarity of cases using features listed in Table 2 and two other important factors: time (temporal) and spatial (location). We shall detail the operations in each level of the 4Rs in the following discussion.

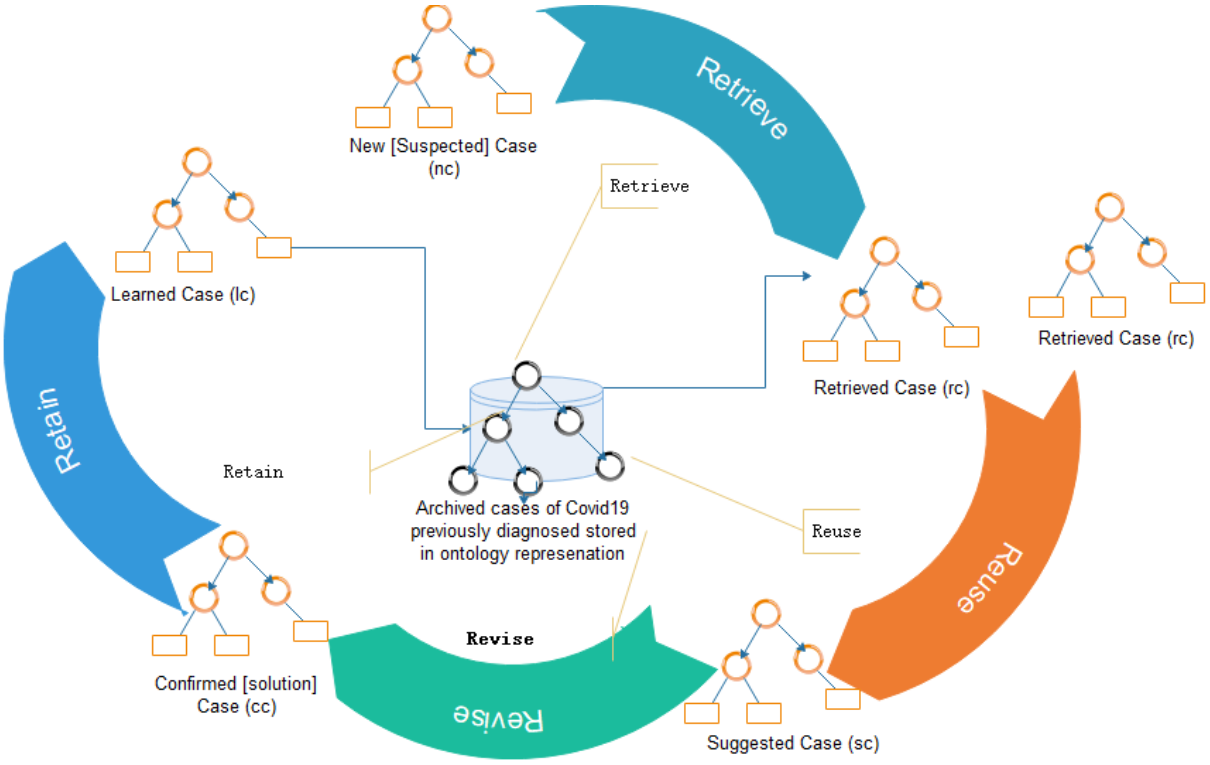

Figure 5: A model of the proposed case-based reasoning (CBR) used in the NLP-Ontology oriented Spatial temporal framework for detecting COVID-19

A. Retrieve

Based on the general concept of the CBR paradigm, the RETRIEVE procedure/algorithm simply uses some efficient distance or similarity computation models like the Euclidean distance, Cosine Similarity [48], and Manhattan distance. Our approach for the procedure of the RETRIEVE algorithm is described as follows: Consider new case *nc* and an archive of stored cases in the CBR knowledge base $SC = \{sc_1, sc_2, sc_3 \ldots sc_n\}$ such that the CBR model RETRIEVE the most similar sc_i or some sc_i from SC. However, the process of retrieval of some sc_i depends on Equation 1. The smaller the value of $Sim(nc, sc_i)$ the more acceptable the case sc_i becomes for adoption for REUSE. Here is a summary of procedures in the RETRIEVE step:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

i. Query generator and parser are used to construct a query that will fetch all similar cases from the case archive SC. The query(ies) is generated based on the extracted features in the previous stage of the framework described in Figure 8.

ii. Semantic Query Web Rule Language (SQWRL) (details later) is employed for modeling the constructed query in the preceding step.

iii. Output resulting from the SQWRL query is sorted in the order of the most similar to the least similar cases. Cases are assumed to be similar if their measure of look alikeness (based on the corresponding features) is non-negligible. The smaller the value of the similarity, the higher the likelihood of the new case (*nc*) to share close similarity with a sc_i or some sc_i, while the bigger the value of the similarity/distance metric, the lower its tendency to match up with *nc*.

iv. Hence, our problem can therefore be modeled as a classification problem whereby some $sc_i \in$ SC are classified to share some similarity with *nc* while another class of some $sc_i \in$ SC are categorized among dissimilar cases. Execution of clinical similarity of cases is done by steadying the computation within the range of [0, 1] using the following equations:

- Euclidean distance: Describes the length between two points and is the most used distance/similarity metric with most appropriate for cases with continuous or dense data. Equation 1 models Euclidean distance.

$$ED = \sqrt{\sum_k^n w_k \,.\, (f_{ak} - f_{bk})^2} \quad (1)$$

- Cosine Similarity: This similarity metric measures the dot product of the two features compared. Based on the cosine computation which yields 1 for 0^0 and less than 1 for other degrees, it implies that a cosine similarity of 1 signals that features A and B are similar cases while a cosine value of -1 indicates non-similarity. Equation 2 models Cosine similarity which has strong application in data with sparse vectors. In addition, the Cosine similarity (CS) is able to perform that Euclidean distance (ED) in cases where ED sees two cases to be distantly similar; CS might observe a closer similarity among the two cases based on their oriented closeness.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

$$CSim(x,y) = \frac{x \,.\, y}{||\mathbf{x}||\,||\mathbf{y}||} \quad (2)$$

where $||\boldsymbol{x}||$ and $||\boldsymbol{y}||$ are the Euclidean norm of vector $x=(x_1,x_2,\ldots,x_p)$ and Euclidean norm of vector $y=(y_1,y_2,\ldots,y_p)$, respectively, and vector x defined as $\sqrt{x_2^1 + x_2^2 + \cdots + x_p^2}$.

- Manhattan distance: Another similarity or distance metric, also known as Manhattan length measures distance between points along an axis at a right angle. Equation 3 models Manhattan distance.

$$MD = |x_{ak} - x_{bk}| + |y_{ak} - y_{bk}| \quad (3)$$

- Other similarity measures are the Jaccard similarity (use for sets) and Minkowski distance equations.

Now, because our cases in the proposed framework were modeled/formalised in ontology represenation, there was a need to be able to carry out quantitative measures of similarity between features of cases, hence the need to use ontology-based semantic similarity between terms. There are six (6) major techniques for computing such similarities of features in ontology: ontology hierarchy approach, information content, semantic distance, approach based on properties of features, approach using ontology hierarchy, and hybrid methods [48]. Therefore, to compare two cases, we make the following assumptions:

i. Two cases are similar if their ontologies demonstrate similarities in both feature values and structure (of their ontological representations).

ii. That an arbitrary weight $\boldsymbol{w_i}$ value is assigned to each property (object and data properties) in their (cases to compare) ontologies (which must sum to 1 for each ontology representing a case). For example, all properties/relations/links (denoting object properties) as shown in Figure 11 from the Case node to the other nodes (Diagnoses Case, Symptoms, Epidemiology, Radiology/Laboratory, Clinical Diagnoses, and Treatment) each will be assigned a weighted value. Similarly, links (data properties) from the second level (Diagnoses Case, Symptoms, Epidemiology, Radiology/Laboratory, Clinical Diagnoses, and Treatment) to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

the lower level (values) also have weight values summing up to 1. For instance, the **presentsSymptom** (object property) may have weight 0.3, **hasEpidemiology** (object property) may have such that each symptom weight 0.2, and so on until all second level nodes have weights summing up to 1. However, a case may present ***n*** features associated with the **presentsSymptom** link, i.e a case with cough, fever, Anosmia and ARDs (which are all symptoms). So, we also assign weights $\boldsymbol{fw_i}$ to each of the known symptoms such that only the weight of the present symptoms in an arbitrary case [***c***] are summed up all $\boldsymbol{fw_i}$ for such case and multiply it with its second level (Symptom node) weight $\boldsymbol{w_i}$. Hence, each mention/use of the notation $\boldsymbol{W_i}$ in our Equation 6 will denote summation of all ($\boldsymbol{w_i}$. $\boldsymbol{fw_i}$)

Note: all our objects (such as **presentsSymptom** and **hasEpidemiology**) and data type properties are detailed and discussed in Section 4.3

iii. We modeled all features in Table 2 according to their expected inputs. For example, some features are either present or absent and those features with values bounded between a range (e.g. Cough: Y/N, Oxygen therapy: Y/N, Heart rate: >=88.63, Temperature (°C): >= 37.86) are modeled with 1 or 0 and values from [0.0, 1.0] respectively. Hence, each .$\boldsymbol{fw_i}$ described in (ii) above affected by their impact is denoted by ***p***. Hence, the true value of .$\boldsymbol{fw_i}$ is :

$$\boldsymbol{fw_i} = . \boldsymbol{fw_i} . \boldsymbol{p}$$

Now that we have established our distance/similarity functions and basoc assumptions for case retrieval, here is the formula for computing similar cases in the archived compared to new case (*nc*). This study adopts the approach based on properties and features described in [49]. The adapted similarity measure is that of Tversky [70] as shown in Equation 4:

$$Sim(nc|sc_i) = \frac{|D_1 + D_2|}{|D_1 \cap D_2| + \mu |D_1 / D_2| + (\mu - 1)| D_1 / D_2|} \quad (5)$$

The model in equation (5) assumes that *nc* and sc_i are cases whose features are collected in D_1 and D_2 respectively. Therefore, the similarity between *nc* and sc_i is computed using three components of equation (5): distinct features of *nc* to sc_i, distinct features of sc_i to *nc*, and common features of *nc* and sc_i, for 0 <= μ <= 1, a function that defines the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

relative importance of the non-common features. D1 and D2 represent the target and the base respectively while || stands for the cardinality of set. Although we approve of the similarity model in Eq. 5, we however saw its limitation which is based omission of the effect of weight on selected features. Therefore, we modified Eq. 5 so that we do not use the elements of the set alone, but the weight-value of the elements in each D_i which is computed by Eq. 6. Hence, our modification to Eq. 5 is shown in Eq. 7, afterward, the most similar sc_i is RETRIEVE and forwarded to REUSE after applying Eq. 8.

Where D_1 or D_2 computed using Eq. 6:

$$D = \sum_{i-0}^{n} \boldsymbol{wi} \,.\, \boldsymbol{fwi} \tag{6}$$

$$Sim(nc|sc_i) = \frac{D_1 + D_2}{D_2 \ \cap \ D_1 \ + \ \mu \, D_1 \ / D_2 \ + \ (\mu - 1) \, D_2 \ / D_1} \tag{7}$$

Furthermore, since $Sim(nc|sc_i)$represents our similarity between new case (***nc***) and an arbitrary case sc_i in the archived, we can compute the similarity score (***SS***) using Eq. 8.

$$Similarity\ Score\ (SS) = T - \ Sim(nc|sc_i) \tag{8}$$

Where T is pre-computed threshold value representing a maximum summation of all possible features a case can have.

Hence, cases with with ***SS*** close to 1 are similar to ***nc*** and as result such cases are retrieved. However, if no case is retrieved by Eqs. 7 and 8, we then conclude that ***nc*** might not have any similar case.

We further compute ***SS*** for positive and negative cases and apply Equations 9 and 10 to determine the following: When $SS_{\text{covid19+}} > SS_{\text{covid19-}}$ the case is classified as positive case of COVID-19, while if $SS_{\text{covid19+}} < SS_{\text{covid19-}}$ the case is concluded to be negative. However, an evaluation of $SS_{\text{covid19+}} = SS_{\text{covid19-}}$ indicates an inclonlusive diagnoses, therefore necessitating more similar case(s) be retrieved.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

$$SS_{\text{covid19+}} = \sum_{k=0}^{n} (SS^{+}) \tag{9}$$

$$SS_{\text{covid19-}} = \sum_{k=0}^{m} (SS^{-}) \tag{10}$$

B. Reuse

The REUSE procedure allows the system to modify the RETRIEVE cases $\boldsymbol{sc_i}$ in such a manner that we have only one similar case. The similar case is constructed to maintain a similar ontology structure with the ***nc*** case. This is achieved by rebuilding an anonymous case (***ac***) by extracting all similar features of the presented cases in sc_i until ***ac*** assumes the form of ***nc***. As such the modified ***ac*** is presented as a temporary solution to ***nc***. The approach proposed here is different from methods used by Gu, et al. [50], which relied on clinical protocols guidelines and medical experts respectively. The ***ac*** case is therefore considered a solved case which will be passed on to the REVISE step for processing.

C. Revise

The evaluation of ***ac*** case at this stage is achieved by ensuring that the summation of case features of the proposed solution case is not greater than 1. If they evaluate to more than 1, some non-essential features are dropped and the weights of the features are recomputed until an appropriate value is obtained. The revised and evaluated case now becomes a candidate case for use, and it is called the repaired case (***rc***). Furthermore, ***rc*** is then used to solve the new problem ***nc*** presented to the system. The solution to ***nc*** is passed to the RETAIN.

D. Retain

Finally, the RETAIN procedure simply stores the solution to ***nc*** as a case that has been learned and is fit to be stored/added to the knowledge base of CBR model for future use.

3.5 Algorithm for Case Retrieval

Algorithm 1 details the complete procedure outlined in proceeding subsections, and describes how a new case of COVID-19 is classified as a positive or negative case using the CBR method. The input to the algorithm is an HER of the new case and the out is Diagnoses Case (Suspected, Confirm, Presumed status).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

Algorithm 1: An algorithm using NLP-Ontology CBR framework for detecting and diagnosing COVID-19

Input: $case - EHR\ (cEHR),\ archived\ cases(ACs)$

Output: Diagnoses Case (Confirm Covid19^{+} | Confirm Covid19^{-})

```
Start
      mtdk [][], rtcf[][], rfb [][]
      sc []
      ac ← ∅
      SFs ←sentenceTokenizer(cEHR)
      SFs← spellCorrector(SFs)
      SFs ←  lexicalNormalizer(SFs)
      SFs ←  sentenceNormalizer(SFs)
     ncsat ←  NSC(SFs)

      mtdk, rtcf, rfb ← ∅
      mtdk ←MTDK(ncsat)
     rtcf ←RTCF(mtdk)
     rfb ← RFB(rtcf)
     nc ← formalizeFeatures(rfb)
     sc ← RETRIEVE(nc, ACs)
     ac ← REUSE(nc, sc)
     rc ← REVISE(ac)
     adapt rc for nc
     ACs +← RETAIN()
     diagnosis_status ←find status of rc
     return  diagnosis_status
end
```

In addition, Algorithm 2 outlines the procedures for ontology learning/population to formalize suspected cases which are presented in natural language representation. The procedures described in Algorithm 2 were defined at a high-level in Algorithm 1. Whereas Algorithm 1 describes a flow of data in the framework in Figure 1, Algorithm 2 however details the task of translating raw text in natural language (English) into ontology formalism (ontology learning). The task of ontology learning here is simply to learn terms/concepts and their instances from raw natural language text. The learned concepts are encoded as terminology box (Tbox) whiles their instances and assertions (class and object) are encoded in the assertion box (Abox). Although Section 4.3 describes the domain ontology (largely the Tbox) engineered in this study, we however note that does not include formalization unknown suspected cases of Covid19 which the framework needs to translates into a feature-based representation.

Algorithm 2: An algorithm detailing the ontology learning/population approach for case-to-feature representation

Input: $raw - text\ of\ suspected\ case$

Output: ontology representation of *of suspected case*

```
Start
     alignment ← ∅;   tboxitems ← ∅; case-ontology-graph ← ∅
     sentences  ← textblob(raw-text)
```

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

```
4     noun_phrase ←EXTRACT(sentences)
5     for sentence ∈ sentences, token ∈ noun_phrase do
5        if EXTRACT_POS_TAG(token) ∈ {NN,NNP,JJ,VBG,RB,CD} then
6           ner ← EXTRACT_NAMED_ENTITY(sentence)
7           syns  ← WORDNET(token, POS)
8           if POS(token)∈{NN,NNP,JJ,VBG} then
9              if COMPARE(token, ner, syns) then  # rule to match if 'token' is concept
10                tboxitems +← token
11             else
12                aboxitem ← token
13                  concept ←EXTRACT_CONCEPT(aboxitem, tboxitems)
14                 property ←EXTRACT_PROPERTY(sentence, token)
15                  range ←EXTRACT_VALUES(ner, concept)
16                case-ontology-graph +← assert(concept, property, range)
17             end
18          end
19       end
20    end-for
21    return case-ontology-graph
22 end
```

4.0 Experimentation

In this section the clinical data and experimentation environment used in this study are described. In addition, we develop the domain ontology (for Covi19 and other related Covs-based disease) and also the case-based ontology for new cases. Finally, we demonstrate the implementation of the framework as shown in Figure 1.

4.1 Clinical Data

The COVID-19 pandemic is currently a global emergency with limited access to health facilities and even computerized patient records which could have allowed access datasets for computational research. Although there are statistical-based datasets accessible in the forms of structured, semi-structured, and unstructured (e.g. WHO, Johns Hopkins University, mainstream news media, and even social media), however, such datasets are still unfit for tasks like the one in this study. After a thorough search for publicly available patient HER-based benchmarked datasets of COVID-19 with none accessible, we decided to adopt the approach of curating new datasets of COVID-19 from some available data on standard domains.

The data curated was obtained from the Italian Society of Medical and Interventional Radiology (SIRM). SIRM is a scientific association which includes the majority of Italian radiologists, and is targeted to encourage the progression of diagnostic imaging by promoting studies and research. The data source (https://www.sirm.org/en/italian-society-of-medical-and-interventional-radiology/) listed English-like records (itemizing age, symptoms and signs manifested, and other laboratory details) and CT scans for each of the sixty-eight (68) COVID-19 patients. We anonymized and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

cleaned the datasets where necessary and extracted the necessary information, storing them in a format appropriate for this study. Figure 6 shows snapshots of a randomly selected case.

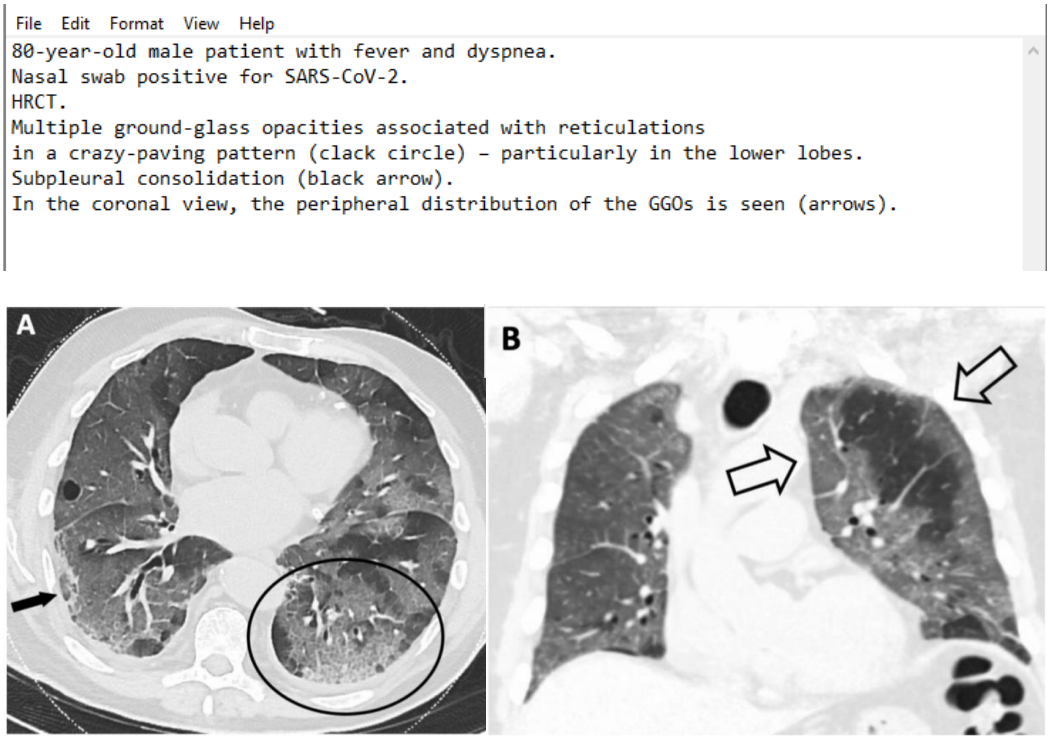

Figure 6: Dataset of a sample case of COVID-19 showing some English-like statements extracted and samples of CT scans and X-ray performed on the patient.

The data need of this study is EHR-based datasets in natural language (NL) format. Hence, we focused on processing the English-like statements extracted for each patient leaving the image-based for future study using the approach of deep learning for classification of COVID-19 cases.

A careful examination of the curated datasets revealed that only 3 cases (case numbers 39, 51, and 60) were confirmed to be negative. However, 47 cases (case numbers 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 29, 30, 31, 32, 34, 35, 36, 37, 40, 41, 42, 43, 44, 46, 48, 49, 50, 52, 55, 56, 61, 63, 64, 65, and 68) were confirmed to be positive, and while 18 cases (case numbers 6, 13, 14, 15, 17, 28, 33, 38, 45, 47, 53, 54, 57, 58, 59, 62, 66 (recovered), and 67) presents inconclusive results due some reasons/events surrounding such cases (e.g death of patient before the conclusion of examination, patient recovered from ailment suspected to be pneumonia, and other unrelated events). We therefore modeled the 3 negative and 47 positive cases accordingly in the archive of the CBR-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

engine. These archived cases form the database from which similarity models and retrievals operations are applied. The 18 unconfirmed cases became a batch of cases from which we drew our input for our framework. Furthermore, we normalized medical records of the positive and negative cases, removing the diagnosis made by physicians and passed each of them as input into the proposed CBR framework. This allows for subjecting our system to the same examination carried out by the experts to establish a basis for comparing the performance of the proposed CBR framework.

4.2 Computational Environment Setup

The implementation was on a personal computer with CPU of Intel (R) Core i5-4210U CPU 1.70 GHz, 2.40GHz; RAM of 8 GB; Windows 10 OS. Furthermore, we deployed Anaconda shipped with Python 3.7.3, SPYDER 3.3.6, and also installed NetBeans IDE version 8.1. The Python platform allows for the implementation of the NLP feature extraction pipeline shown in Figure 2 while the NetBeans IDE provides support for implementing the feature to ontology representation and also the CBR-engine. Modeling of ontologies in this study was achieved using Protégé (Protégé).

4.3 Domain Ontology Modeling

Ontologies are formalism for specification of concepts or abstract description of a system in a domain-specific knowledge composition. Ontologies as formalism stems from description logic (DL) and with support for reasoning it has received OR has caused it to receive more and more attention in computational biology and bioinformatics. There are different ontology languages like RDF/RDFS, DALM+OIL, and OWL. OWL is a DL-based ontology language with high expressivity and has three variants: OWL-DL, OWL-full and OWL-lite. This study models ontologies using OWL2 [51, 52], which is an improved version of OWL (sometimes known as OWL1). We have modeled three different ontologies: the first represents domain knowledge, the second is a formalism of the archived cases, and the third ontology formalizes new cases.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

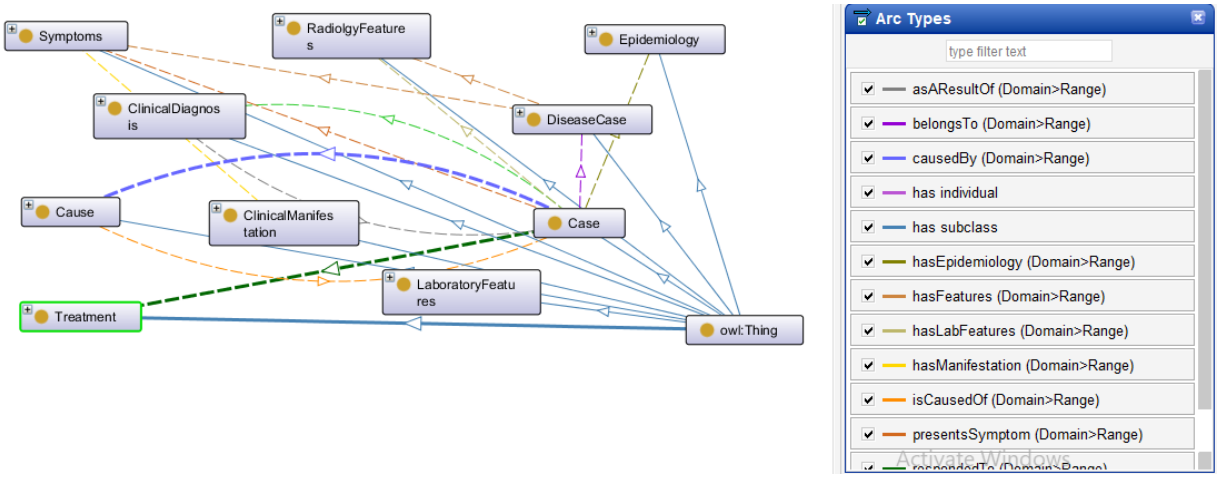

Figure 7: Ontology representation domain-based mapped tokens (features) of a new case of coronavirus (COVID-19): New Case (nc)

In Figure 8, we show a visualization of the ontology representing a new case (***nc***). The ontology captures concepts/classes like ***Symptoms***, ***ClinicalDiagnosis***, ***ClinicalManifestation***, ***Epidemiology***, ***RadiologyFeatures***, ***LaboratoryFeatures Case*** (to denote a new case), ***DiseaseCase***, ***Cause*** (to capture the likely causes of the disease in a case), and ***Treatment*** (which represents treatments administered to a case). Each of these concepts is related/linked to another concept by a property (object property) with almost all concepts linked to **Case**. To the right is a list of the object properties. For example, the line connecting ***Case*** to ***Symptoms*** is the object property ***presentSymptom***. The ***Case*** is the domain while ***Symptoms*** is the range for the object property ***presentSymptom***. Some concepts have the + symbol at the top-leftmost corner of their bounding boxes. This is an indication that there are other subclasses in that concept/class which can be revealed by clicking on the + symbol.

Case formalization is therefore made possible through the Case-Based ontology file shown in Figure 8. While that illustrates a case of COVID-19, we made a further effort to use an ontology approach to model the archive of stored cases in the CBR engine. To archive this, we represented the structure and the semantic of the information content of such archive using the ontology visualized in Figure 9. As mentioned earlier, the ontology file was modeled and visualized in Protégé (Protégé)**.** The ontology consisted of 459 axioms, 225 logical axioms, 213 declaration axioms, 196 Class, 11 object property, 8 data type property, 181 subclasses, and 15 instances (with the exception of cases of COVID-19 which forms the archive of cases in the ontology). Figure 9 captures the ***is-a*** relationship existing among

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

classes and Figure 10 outlines all the classes (and their hidden subclasses), instances (individuals) of the declared classes, object and data type properties.

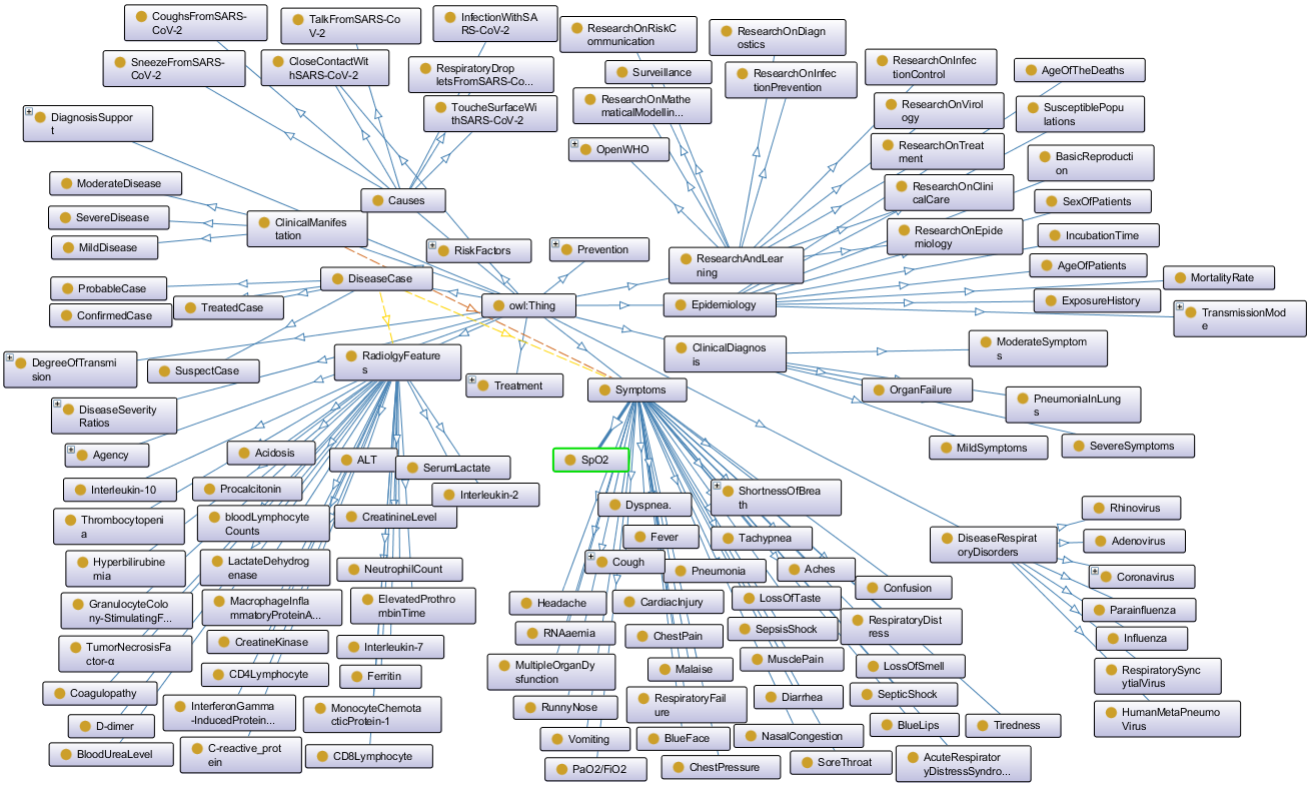

Figure 8: A visualization of ontology representation of relations of concepts (TBox) in a domain-based knowledge repository of COVID-19 using the ***is-a*** relationship

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

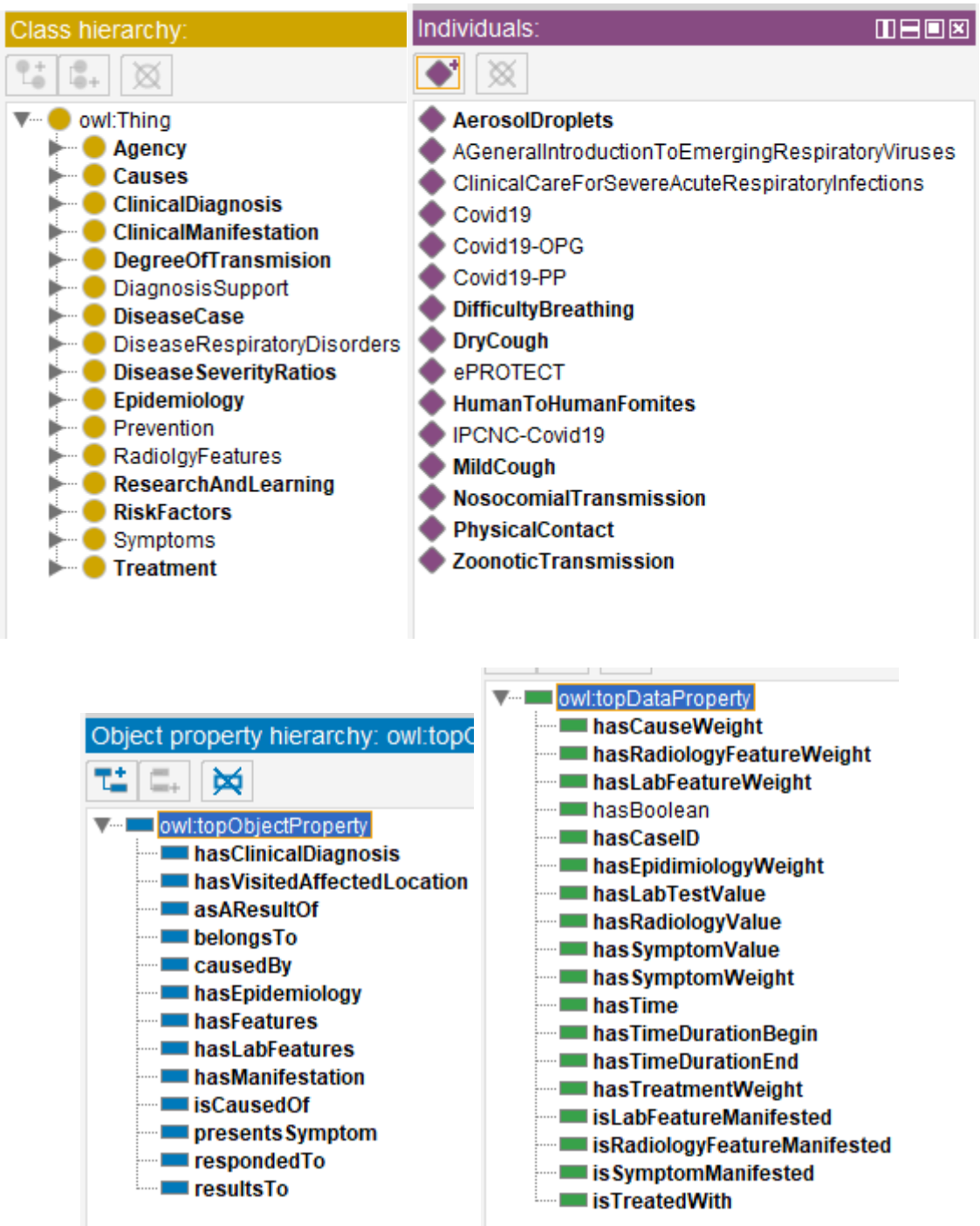

Figure 9: A listing of concepts (classes), individuals (instances of classes), object properties and data type properties modeled in a domain-based knowledge repository of COVID-19

Now that we have formalism for archiving all cases in the proposed framework and also a formalism for modeling new cases extracted from electronic records, we shall consider how the proposed approach will implement its query

 doi:10.20944/preprints202005.0171.v2

for similar cases as modeled using mathematical models in item A of subsection 3.4. To archive an optimized and effective query of cases from the archive, we decided to construct our query from the mathematical models presented earlier using **S**emantic **Q**uery-Enhanced **W**eb **R**ule **L**anguage (SQWRL), pronounced *squirrel*. SQWRL is a query language primitive to OWL and also an SWRL-based with syntax of SQL-like and having operators for extracting information from OWL ontologies [56, 57]. We chose SQWRL over SPARQL because of its suitability for use in OWL ontologies since it does not require serializing our OWL ontologies in RDF/RDFS, an operation which often causes a knowledge-base (ontology) to lose some semantics and expressivity as a result of serialization. Moreover, the rule-form of SQWRL and its compatibility with the rule language SWRL allows for improving our framework to use inference engine, thereby improving the knowledge-base through inference. Protégé also provides a tab for executing our SQWRL queries against the ontology through the SQWRLTab plugging. We may take advantage of this tab to text our generated queries, although the framework proposed in this study has a mechanism for doing the query execution automatically through OWLAPI.

Now for instance, given a new case (***nc***) presenting with the following features according to their category, we might be interested in translating our mathematical model into an SQWRL query such that similar cases are retrieved: *Symptoms (Cough, Temperature, Nausea and vomiting, Shortness of breath, Contact with case(s)); Laboratory Features(Neutrophil, Lymphocyte, Active partial thrombin time);*

The following conditions can be assumed from our case: retrieve all cases according to their value of similarity (in descending order), which have values for all or some of the features: *Symptom (Cough, Temperature, Nausea and vomiting, Shortness of breath, Contact with case(s)); Laboratory Features (Neutrophil, Lymphocyte, Active partial thrombin time).* In Figure 17, we present a sample SQWRL-query for extracting similar cases compared to the new case example described here.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

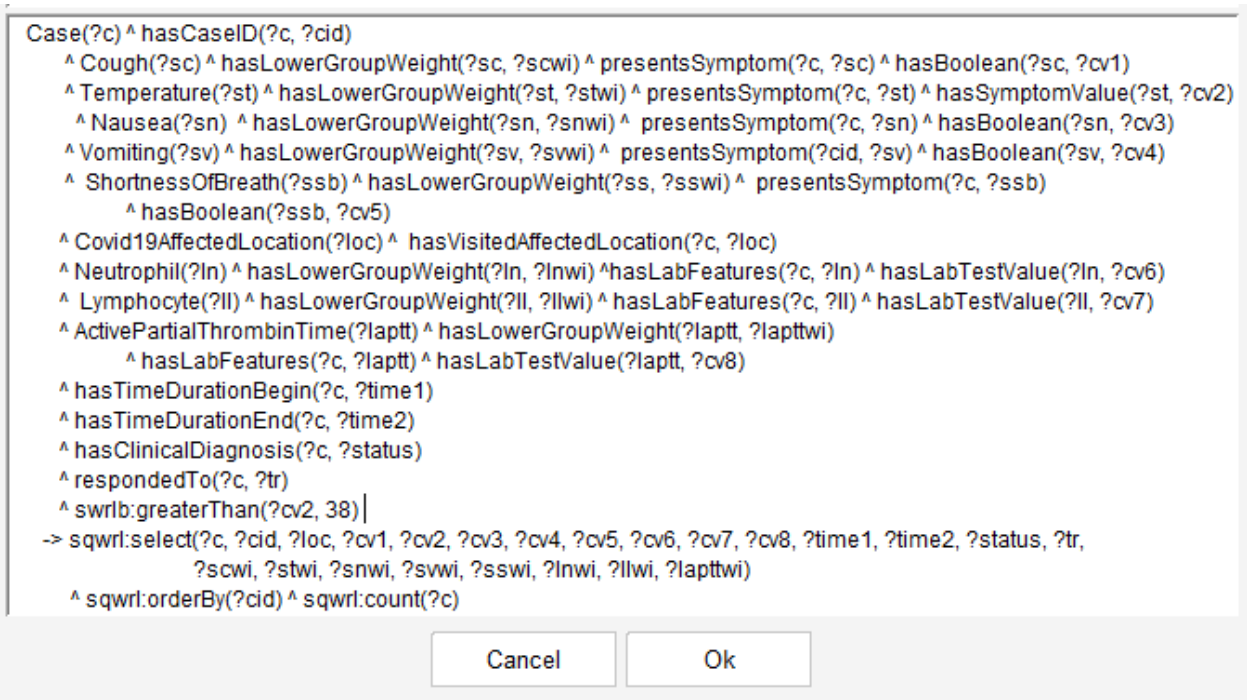

Figure 10: A sample SQWRL query constructed to retrieve similar cases corresponding with the new case (*nc*) accepted as input.

The sample query in Figure 10 was submitted to the Protégé application for execution and query of the underlying ontology through the SQWRLTab plugins. The syntax of the SQWRL query language aligns itself to the declared classes or entities, properties (both data type and object), and instances/individuals on the ontology. This is why you will observe that the predicates (unary and binary) names in the listed query in Figure 17 derive their values from the declared classes or entities, properties (both data type and object), and instances in the ontology. This positions the SQWRL query above the use of SPARQL. A detail explanation of the query given in the following lines:

The first line **`Case(?c) ^ hasCaseID(?c, ?cid)`** extracts all cases and their case IDs from the CBR case archive and stores those two values in the **`?c`** and **`?cid`** variables. Furthermore, the second lines 2-5 of our query select instances of the following symptoms which were keywords/features extracted from the natural language input above: **`Cough`**, **`Temperature`**, **`Vomiting`**, and **`ShortnessOfBreath`**, and their weight values. This is summarized in the following lines:

```
          ^ Cough(?sc) ^ hasLowerGroupWeight(?sc, ?scwi) ^ presentsSymptom(?c, ?sc) ^
hasBoolean(?sc, ?cv1)
          ^ Temperature(?st) ^ hasLowerGroupWeight(?st, ?stwi) ^ presentsSymptom(?c, ?st) ^
hasSymptomValue(?st, ?cv2)              ^ Nausea(?sn)  ^ hasLowerGroupWeight(?sn, ?snwi) ^
presentsSymptom(?c, ?sn) ^ hasBoolean(?sn, ?cv3)
          ^ Vomiting(?sv) ^ hasLowerGroupWeight(?sv, ?svwi) ^  presentsSymptom(?cid, ?sv) ^
hasBoolean(?sv, ?cv4)
```

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

```
^  ShortnessOfBreath(?ssb) ^ hasLowerGroupWeight(?ss, ?sswi) ^  presentsSymptom(?c,
?ssb) ^ hasBoolean(?ssb, ?cv5)
```

Also, our natural language based query has some laboratory features which we also extracted their values for each of the cases retrieved due to the query on line 1. These laboratory features are queried as follows:

```
^ Neutrophil(?ln) ^ hasLowerGroupWeight(?ln, ?lnwi) ^hasLabFeatures(?c, ?ln) ^
hasLabTestValue(?ln, ?cv6)
^  Lymphocyte(?ll) ^ hasLowerGroupWeight(?ll, ?llwi) ^ hasLabFeatures(?c, ?ll) ^
hasLabTestValue(?ll, ?cv7)
^ ActivePartialThrombinTime(?laptt) ^ hasLowerGroupWeight(?laptt, ?lapttwi)
^ hasLabFeatures(?c, ?laptt) ^ hasLabTestValue(?laptt, ?cv8)
```

We are also interested in the probable points/places of contacts/visited by the cases that have been retrieved so far. Hence the line below:

```
^ Covid19AffectedLocation(?loc) ^  hasVisitedAffectedLocation(?c, ?loc)
```

Now that all existing cases in the archive satisfying the above conditions have been retrieved, we further limit the cases to be extracted to the conditions below:

```
^ hasTimeDurationBegin(?c, ?time1)
^ hasTimeDurationEnd(?c, ?time2)
^ hasClinicalDiagnosis(?c, ?status)
^ respondedTo(?c, ?tr)
```

The first and second lines simply ensure the cases retrieved have the time/date when the case manifested and either died or recovered. Line three also allows each case to fetch the result of its clinical diagnosis (Positive or Negative diagnosis). Finally, the `respondedTo(?c, ?tr)` predicate fetches the treatment (if any) options recorded against each case.

Once all these cases are matched by the rule-like left-hand-side (LHS) of our query (a simulated of semantic web rule langue SWRL), the right-hand-side (RHS) uses the `sqwrl:select` predicate to fetch all cases (and their attributes/features) satisfied by LHS using the variables. Hence the lines below:

```
-> sqwrl:select(?c, ?cid, ?loc, ?cv1, ?cv2, ?cv3, ?cv4, ?cv5, ?cv6, ?cv7, ?cv8, ?time1,
?time2, ?status, ?tr,?scwi, ?stwi, ?snwi, ?svwi, ?sswi, ?lnwi, ?llwi, ?lapttwi)
```

Finally, we are interested in counting the number of cases retrieved after ordering them according to their case IDs. The line of query below does this:

```
^ sqwrl:orderBy(?cid) ^ sqwrl:count(?c)
```

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

All cases retrieved by the sample query above must have its features represented in the ontology for the query to be able to match them. Case representation is covered in Section 3 of this paper, however, we have captured in Figure 9, a formalization of sample patient record shown in Figure 6. The case representation shown here is a Protégé interface format of the case, although the ontology notational is equally generated.

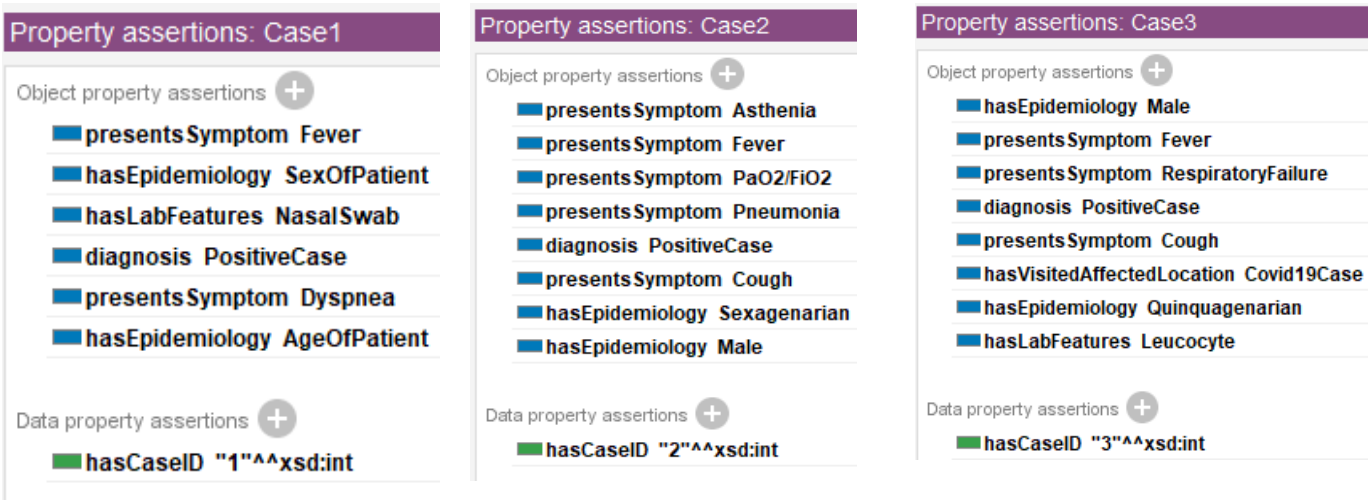

Figure 11: An illustration of case representation as shown in Protégé for cases a 1, 2, and 3 from the 68 cases extracted from the data source

4.4 Implementation and Experiments

The implementation of the CBR framework proposed in this study adopted JCOLIBRI [55]. Jcolibri is a library containing APIs for implementing a CBR framework and is written in Java. As a result, we employed the use of Java programming language to develop the CBR-engine (shown in the right box or component of Figure 12), Python programming to implement the natural language to Normalized Sentence Component (NL-NSC), and finally, a combined use of the two languages made the implementation of the feature extraction and formalization components of Figure 12 possible.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

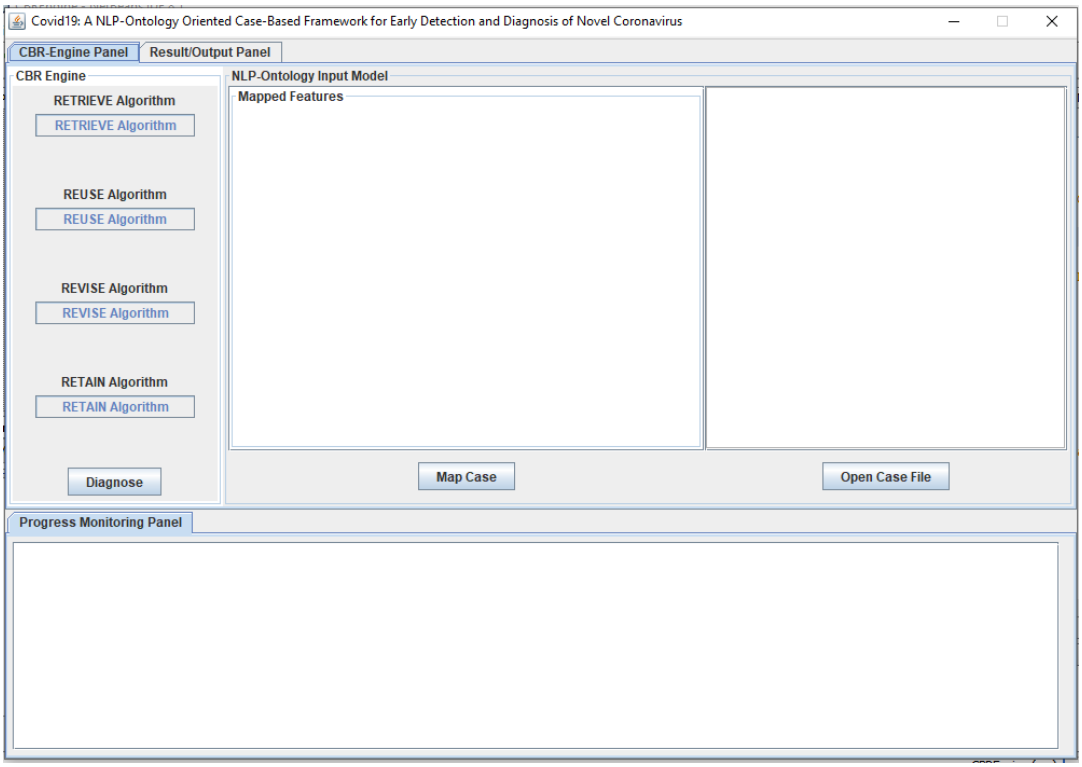

Figure 12: A graphical user interface (GUI) showing the major components of the proposed CBR-based framework for classifying cases of COVID-19 as either positive or negative case.

The complete implementation of the proposed framework is accessible through a graphical user interface (GUI) designed for this study and shown in Figure 11. The file loader and raw text extraction component of Figure 9 is implemented in the rightmost panel with a box and 'Open Case File' button in Figure 12. Furthermore, from Figure 8, the center panel containing a box and 'Map Case' button captures the implementation of the NL-NCS, feature extraction, and feature formalization components identifiable from Figure 11. To achieve this, standard Python libraries and NL-based libraries (like NLTK and Stanford CoreNLP) were richly employed to carry out the tasks of sentence disambiguation, spelling correction, lexical normalization, and normalization of sentences into their corresponding structures or components, and tokenization of sentences to enhance the process of feature mapping. However, the feature mapping and formalization of cases in ontology format was achieved using OWLAPI, Wordnet API, and Pellet API (an OWL-based knowledge reasoning plugin) which were implemented through a skillful use of Python and Java.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

The result of the extracted and mapped features presented us with a challenge of accurately extracting values from the processed patient record. For example, we could have extracted features like 'Fever', 'Temperature' and so many other features which largely rely on syntax and semantic parsing of domain lexicon. But the challenge we were faced with was detecting the semantics/meaning and context of usage of the features from the patient records. To circumvent this, we took advantage of the named entity resolution technique we applied to the text.

At a sentential-level, an attempt was made to search for values of features within the neighborhood of that feature. For instance, given the sentence:

'*The temperature of the patent was 38^{oc}'*,

a careful parsing of the sentence using NLP technique will reveal that the feature (temperature) has 38 degrees Celsius. But consider the sentence:

'80-year-old male patient with fever and dyspnea.'

There are two features in the sentence (fever and dyspnea) which do not have explicit declaration of values assigned to them. In cases like these, we developed a sentiment analysis component which enabled us to detect if such features were stated in the affirmative or negative form. The outcome of our sentiment analysis model outputs was: positive, negative and neutral. These outputs were used accordingly to formalize the feature and its value (true or false, as shown in Figure 16 and Table 2) in the ontology.

The leftmost panel of Figure 12 illustrates the implementation of the CBR-engine. This CBR-engine and the mathematical model presented in Section 3 were implemented with Java using the jcolibri API which models the Retrieve, Retain, Revise, and Reuse (4Rs) of CBR paradigm, allowing for users to adapt it to their frameworks. Meanwhile, we have also added a panel for monitoring the procedures for detection of status of any presented case of COVID-19; this monitoring begins with file loading component to the CBR-engine processes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

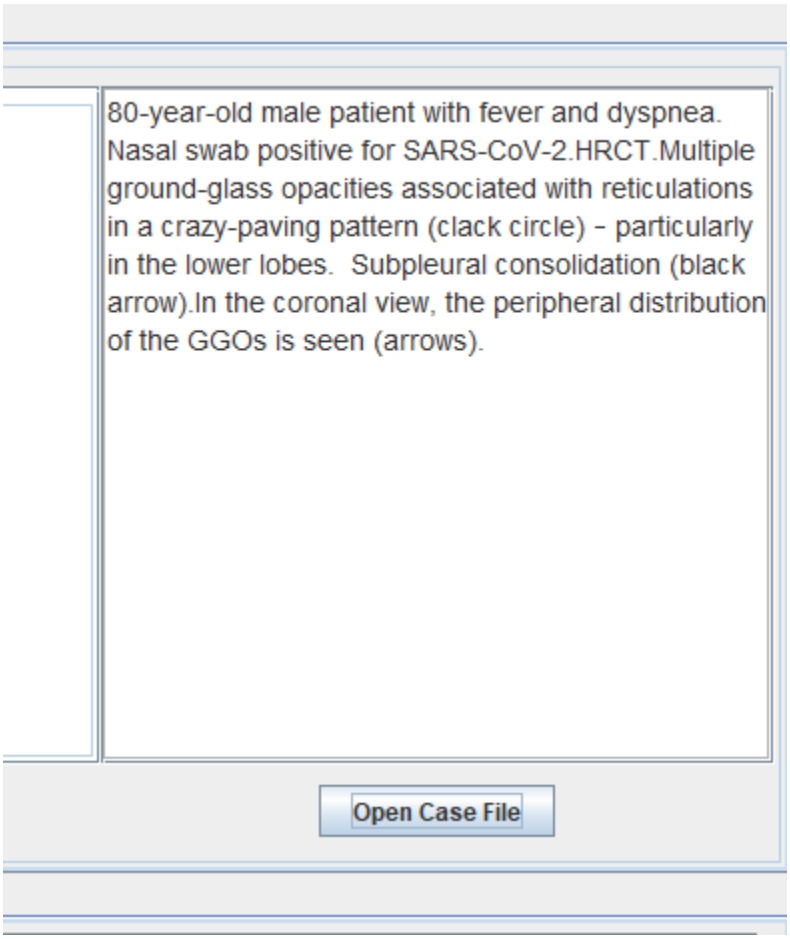

Figure 13: A demonstration of the File Loader component of the proposed framework

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

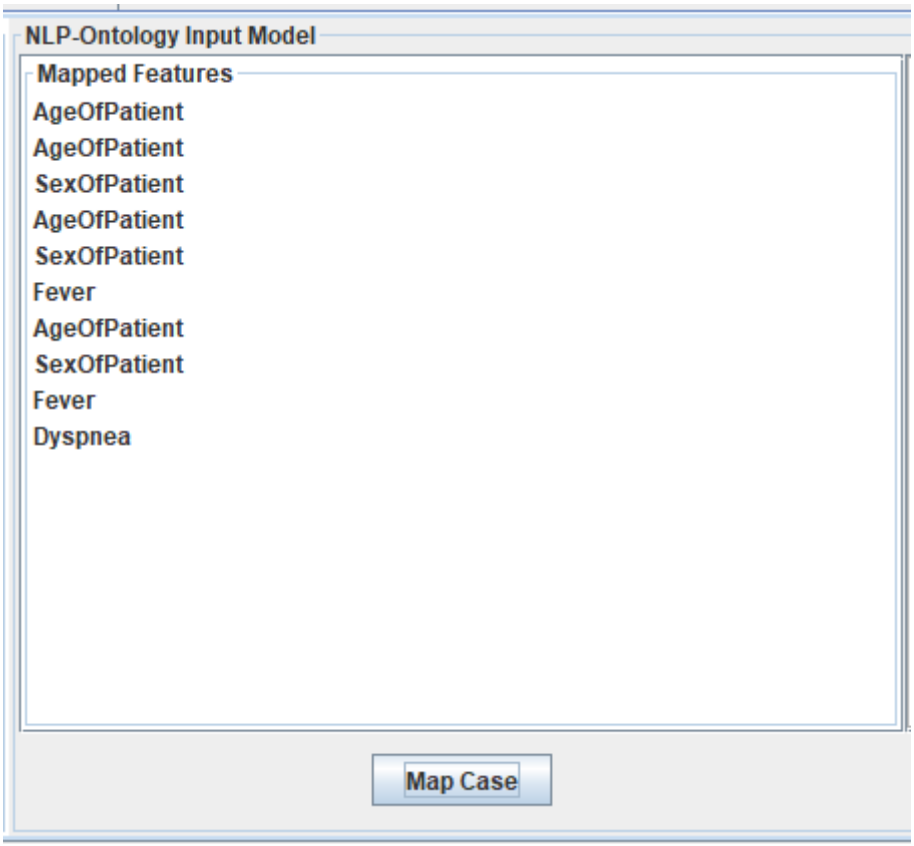

Figure 14: A demonstration of the Feature Mapping component of the proposed framework

Experimentation using the datasets discussed in Section 4.1 revealed that the implementation of the proposed CBR framework was successful. Figures 13 and 14 show a demonstration of the File Loader and Feature Mapping components of the CBR framework. Meanwhile, the process of formalizing feature-value relationship was monitored and is shown in the Progress Monitoring panel in Figure 15. Lines delimited and prefixed by the **<<<Derived>>>** symbol represents components of the generated new case ontology learnt by our Algorithm 2. Each line is an assertion resulting from the features extracted from the input raw-text.

The progress monitoring panel output shown in the Figure 15 demonstrates how Algorithm 2 successfully extracts features from the first sentence of the patient record shown in Figure 13. The output is then further translated into an ontology formalism representing the new case (or new problem) the CBR model receives as input for further extraction similar cases using SQWRL-based query as illustrated in Figure 7.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

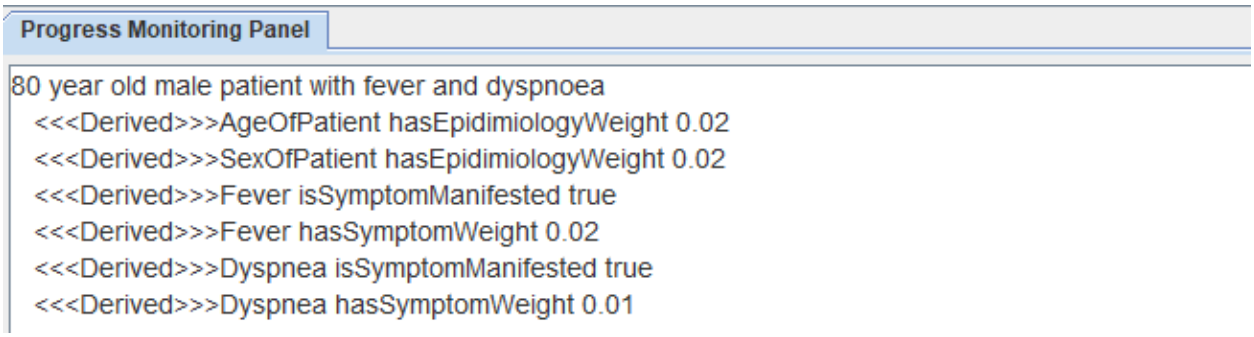

Figure 15: A demonstration of the formalization of feature-value extracted in ontology representation

The CBR-engine then collects the ontology representation of the new case for the purpose of reasoning operation. The task of classification of any suspected case of Covid19 model in Figure 15 now rests on the CBR-engine which is detailed in Section 3.5.

After a complete testing of the implemented framework using our datasets, we discovered that the classification accuracy of the improved CBR model yielded an interesting result as shown in Figure 16.

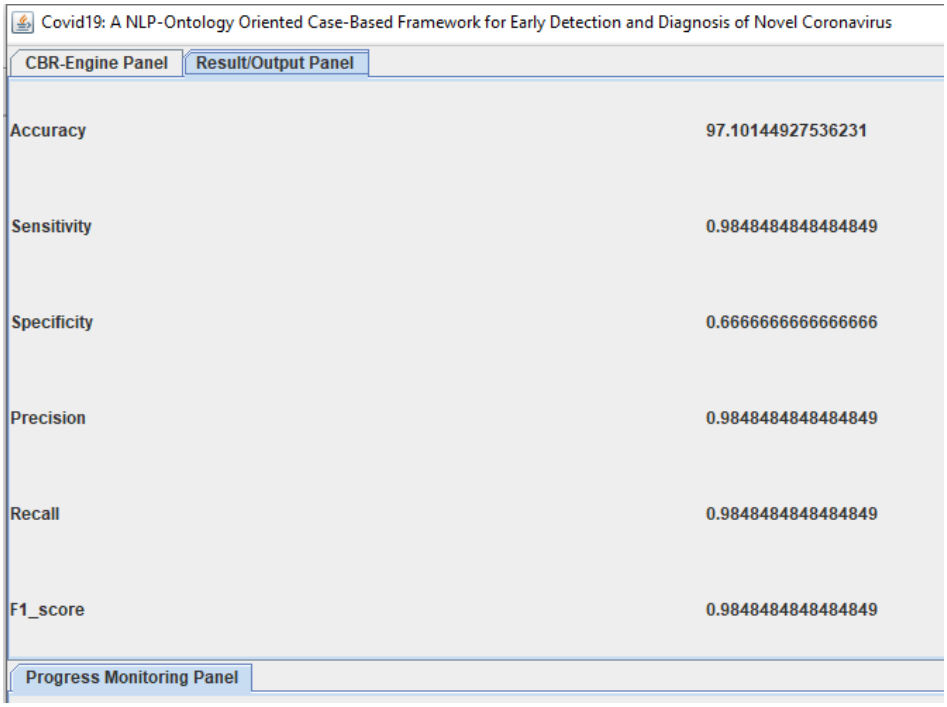

Figure 16: Result of diagnosis of a case of COVID-19 showing the status (positive or negative), clinical diagnosis (acute, mild or severe), estimated duration (in days), and likely treatment

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

5.0 Result and Discussion

In this section, we present the performance of the proposed CBR framework compared to the performances of other similar systems. The following are the metrics and their corresponding formula used in analyzing the performance described in this section.

i. Accuracy= (TP+TN)/(TP+TN+FP+FN)
ii. Specificity= TN/(TN+FP)
iii. Sensitivity= TP/(TP+FN)
iv. Precision= TP/(TP+FP)
v. F1=(2*Recall)/((2*Recall)+FP+FN)
vi. F=(2* Precision * Recall)/(Recall + Precision)
vii. Recall= TP/(TP+FN)

Note that the following are the derivations for the TN, TP, FN, and FP:

TN = Suspected cases of COVID-19 which both the proposed CBR framework and the curated dataset presented concluded to be negative cases of COVID-19.

TP = Suspected cases of COVID-19 which both the proposed CBR framework and the curated dataset presented as being positive with COVID-19.

FN = Suspected cases of COVID-19 which the proposed CBR framework concluded to be negative cases of COVID-19 while the curated dataset presented as being positive with COVID-19.

FP = Suspected cases of COVID-19 which the proposed CBR framework presented as being positive with COVID-19, while the curated dataset shows negative cases of COVID-19.

5.1 Comparison of the ontology of the systems with others

The ontologies developed in the research are very tangible in enhancing the performance of the proposed CBR framework. However, to measure the performance and importance of the knowledge representation formalism used in this study, we resolved to compare the efficiency of the proposed ontology with other related ontologies used in related studies on CBR by using the following metrics:

i. **Class Complexity:** Average number of paths to reach a class from the *Thing* class
ii. **Class Complexity**: Average number of semantic relations for object properties per class
iii. **Abstraction:** Average depth of the ontology
iv. **Cohesion:** Average number of connected classes

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

v. **Semantic richness:** Ratio of total number of semantic relations mapped to classes, by all ontology relations consisting of object properties and subsumption relations.

vi. **Inheritance richness:** Average number of subclasses in a class.

vii. **Attribute richness:** Ratio of total number of data type properties by the number of classes.

viii. **Comprehension of properties (object and data type):** Percentage of annotation of the properties in the ontology

ix. **Comprehension of classes:** Percentage of annotation of the classes in the ontology

Based on these metrics, the performance measurements in the following subsections are presented. Figure 17 shows how the values used in computing the metrics were derived through the Protégé while viewing the ontology.

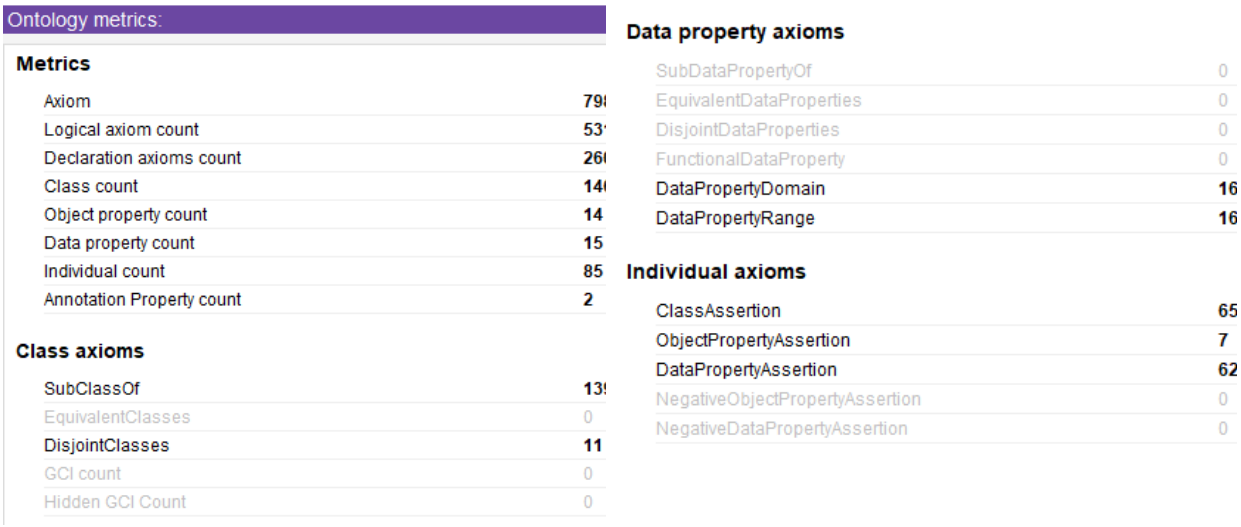

Figure 17: An outline of Protégé-based metrics which yielded values for computation of the comparison metrics used in Tables 3 and 4.

Table 3: An evaluation of some related ontologies used in similar CBR studies in comparison with the proposed ontology as shown in Figures 7 and 8

	Ontology Metrics								
Metrics	**Complexity**		**Abstraction**	**Cohesion**	**Conceptualization richness**			**Comprehension**	
Studies (Year) [Ref]	Class	Properties			Semantic	Data property	Inheritance	Classes (%)	Property (%)
Proposed framework	4	2	4	181	1.041	0.04	8.2	10	10
El-Sappagh and Elmogy [60]	5	1.4	2	63	0.495	2.26	5.0	88.71	2.04%
Heras, et al. [61]	3	1.3	2	27	0.62	0.92	2.875	0.0	0.0

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

The results of Tables 3 and 4 shows the richness of the axioms, properties (object and data type) and instances of the proposed ontology used in this study.

Table 4: An evaluation of some related ontologies based on the contents of their terminology box (Tbox)

	Metrics						
Studies (Year) [Ref]	**Number of individuals**	**Properties with domain/range (%)**	**Number of properties**	**Documentation of properties (%)**	**No. of classes**	**Documentation of the classes (%)**	**No. Axioms**
Proposed ontology	134	100/100	19	10	196	10	1078
El-Sappagh and Elmogy [60]	2640	98.47%/ 98.98%	196	2.04	62	88.71	1316
Heras, et al. [61]	0	85.48/ 77.41	62	0	26	0	446

5.2 Presentation of the accuracy, sensitivity and specificity of the proposed approach

In this section, we present the performance of the proposed CBR model using diagnosis metrics like accuracy, specificity, sensitivity, precision, recall, and F1-score. The choice of these metrics was informed by the peculiarity of the relationship of the values with the disease. For instance, diagnostic accuracy metric was used to evaluate the ability of a diagnostic test to correctly identify a target condition (COVID-19 in this case). This metric particularly is very applicable to cases of diagnoses in medicine since it allows for increased confidence and acceptability of results. In addition, accuracy of diagnosis could help to determine the difference between life and death, so that a system which outperforms another may be seen from an improved accuracy, which also leads to reliability of diagnoses results. Other metric considerations for performance measure in this study were sensitivity and specificity which are also referred to as True Positive Rate (TPR) and True Negative Rate (TNR) respectively. Sensitive and specificity of our system as shown in Table 5, implies the number of COVID-19 cases withthe condition who had a positive result, and the number of COVID-19 cases who did not have the disease and had a negative result respectively. The relationship between these two metrics with respect to accuracy of diagnosis is that the latter allows for the evaluation of the former.

Looking at the results of the metrics above as presented in Table 5, we discovered the proposed CBR framework had a good accuracy. In addition, the sensitivity and specificity value indicated that our system was able to correctly classify cases of COVID-19 as either positive or negative respectively. The precision value means that an average of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

2 COVID-19 cases can be effectively detected by the proposed CBR framework as negative while the remaining 8 out of 10 cases are positive. Similarly, the recall value is 99% which means that approximately 10 out of 10 cases of COVID-19 are correctly classified as positive. F1 score presents the ability of our framework to classify the cases of the disease since the metrics represents a harmonious mean of precision and recall. The precision and recall results therefore portray relevance of positive cases and proportion of correct positive cases are respectively. These are sometimes referred to as Positive Predictive Value (PPV) and True Positive Rate (TPR) respectively.

Table 5: Performance evaluation of the proposed CBR framework using the accuracy, sensitivity, specificity, precision, recall and F1 score

	Performance Metrics					
	Accuracy	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**F1 Score**
Values	97.10%	0.98	0.66	0.984	0.984	0.98

In the next section, we shall compare the performance recorded by the proposed CBR framework with similar studies.

5.3 Comparing the proposed approach with similar methods

A comparative analysis of the performance of our proposed approach was carried out with other case-based reasoning studies. Although the domain of application of the CBR models reviewed differs (medical and non-medical), we discovered that the most important factor lay in the formalism of cases and condition and similarity measures for retrieval of similar cases. An approach of CBR with dominance in the list of studies reviewed using fuzzy logic and those whose cases used ontology for formalism purpose. Also, we observed that some studies investigated the peculiarity and importance of different similarity metrics like Euclidean distance, cosine similarity and others. The effect of such choice of similarity measure helped them to discover the performance effect of a selected metric. The decision of the selection of distance/similarity measure is sometimes influenced by the formalism in which case features are represented. Considering the wide adoption of the use of ontologies as a tool for formalizing cases and its features, we discovered that the interesting performance of this study must have drawn much benefit from the ontology approach for knowledge modeling. An interesting consideration made in this study which makes it outperform other similar works is the choice of a semantic and ontology-based similarity measure metric. We observed that this allowed for a better comparison of cases during retrieval.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2020 doi:10.20944/preprints202005.0171.v2

Furthermore, the novelty of the approach proposed in this study was also uncovered as we compared it with similar studies in the last decade. Only our study adopted the use of NLP technique in extraction of features represented in a presenting case. This allowed for a non-partial automation of the process of diagnosis/detection/classification of cases. We argue that such approach allows for an increase in the level of acceptance of the CBR paradigm. This deduction was made based on the popular manual approach for the extraction of cases and their features from documents represented using natural language. Although some fuzzy-CBR frameworks which were reviewed and compared in Table 6 demonstrate good performance, they are, however, surpassed by our model which combines the techniques of NLP and machine learning (sentiment analysis) in extraction of features in any presenting case. As a result, we presume that an investigation into the hybridization of a CBR model using fuzzy logic, NLP and ontologies may yield a very encouraging performance, and thereby position CBR paradigm as a competitive option for reasoning tasks in artificial intelligence (AI).

In any medical system, the result of diagnosis is more important because the patient has so much to lose when there is a misdiagnosis. So, both under-diagnosis and over-diagnosis are both errors in medical systems and have been a source of concern to wide acceptance for AI-based diagnostic and detection systems in medicine. While over-diagnoses may have to do with over stating the condition of diagnosed case, under-diagnosis is a condition where a diagnosed case does not go on to cause any symptoms or ill-health. This can result in the blurring of the borders between health and disease. Therefore, a diagnostic accuracy helps to investigate how well a particular diagnostic test is able to identify a target condition, in comparison to a reference test. In this study, we carried out our comparison of the proposed CBR framework with other similar studies using diagnostic accuracy. As seen in Table 6, the accuracy of our framework outperforms those of previous studies we compared. Most interesting is the capability of our CBR framework to detect the novel coronavirus (COVID-19) at a higher accuracy. This, therefore, positions this study as a candidate for further improvement of CBR models in future works seeking to diagnose any family of the Covs diseases.

Table 6: A summary of some case-based reasoning (CBR) models and framework, and their domains of application, approaches/techniques used, description of approach and accuracy of the systems

Studies [Ref]	**Year**	Approach used for reasoning or diagnoses	**Domain of Application**	**Accuracy (%)**

 doi:10.20944/preprints202005.0171.v2

Proposed framework	2020	CBR and NLP, and Semantic Web	Detection and diagnosis of COVID-19 (Novel Coronavirus)	97.10
Rahim et al. [61]	2019	Traditional CBR	Diagnosis of psychological disorders	-
Zhong et al. [63]	2018	Text-CBR and ontology	Non-medical: Fault diagnosis and predication by cloud computing	-
Zhang et al. [64]	2017	Traditional CBR	Non-medical: Theory of inventive problem solving for inventive design	-
El-Sappagh and Elmogy [60]	2015	Fuzzy-CBR, and Ontologies	Diabetics	97.67
Shen et al. [65]	2015	CBR with ontology approach	Diagnosis of gastric cancer	-
Heras et al. [61]	2013	CBR with ontology approach	Non-medical: multi-agent systems	-
Li and Ho [66]	2009	CBR and fuzzy logic	Non-medical: Prediction of financial activity	92.36
Petrovic *et. al.* [67]	2011	Traditional CBR	Radiotherapy planning	84.72
Fan et. al. [68]	2009	CBR, Fuzzy decision tree	Medical data classification: breast cancer and liver disorders	98.40 and 81.60
Begum et. al. [69]	2009	CBR and fuzzy logic	Medical data for diagnosis of stress	90.00

6.0 Conclusion

This study is largely focused on adapting a CBR concept to the problem of classifying cases of COVID-19 as either positive or negative even when the disease is in its early stage in the presented case. An NLP model for feature extraction of a presented case was designed and implemented. The innovation of the work presented here lies in sentence-level extraction of feature-value pair for all a-priori declared features. Furthermore, the case retrieval similarity metric applied to the CBR framework proposed in this study contributed to the interesting performance of the system. Meanwhile, knowledge representation (archive of cases stored in the CBR) in the proposed framework was achieved using ontology-based knowledge formalization technique. In addition, new cases were also formalized using ontologies so as to allow for a homogenous basis for case comparisons. The result obtained shows that our proposed framework outperformed state-of-the-art CBR studies with similar approaches to the one in this study. In future, we intend to investigate the performance of our retrieval algorithm over different similarity/distance measure metrics. This will allow for future studies using ontologies and CBR paradigm to effectively select or even combine similarity metrics. Also, we intend to hybridize the proposed method with machine learning methods which allow for application classification algorithm such as SVM.

Declaration of Interests

The authors declare that they have no either financial or personal interests that could have influence the work reported in this paper.

References

1. Naudé, Wim, Artificial Intelligence Against Covid-19: An Early Review. IZA Discussion Paper No. 13110. Available at SSRN: https://ssrn.com/abstract=3568314
2. Pirouz, B., Shaffiee Haghshenas, S., Shaffiee Haghshenas, S. and Piro, P., 2020. Investigating a Serious Challenge in the Sustainable Development Process: Analysis of Confirmed cases of COVID-19 (New Type of Coronavirus) Through a Binary Classification Using Artificial Intelligence and Regression Analysis. Sustainability, 12(6), p.2427.
3. Al-qaness, M.A., Ewees, A.A., Fan, H. and Abd El Aziz, M., 2020. Optimization method for forecasting confirmed cases of COVID-19 in China. Journal of Clinical Medicine, 9(3), p.674.
4. Ting, D.S.W., Carin, L., Dzau, V. and Wong, T.Y., 2020. Digital technology and COVID-19. Nature Medicine, pp.1-3.
5. Vaishya, R., Javaid, M., Khan, I.H. and Haleem, A., 2020. Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews.
6. Leung, G.M. and Leung, K., 2020. Crowdsourcing data to mitigate epidemics. The Lancet Digital Health.
7. oton, P. A. 1988. Using Experience in Learning and Problem Solving. MIT Press.
8. Bareiss, E. R. 1989.Exemplar-Based Knowledge Acquisition: A unified Approach to Concept Representation, Classification and Learning. 300 North Zeeb road, Ann Arbor, MI 48106-1346: UMI.
9. Gierl, L., Bull, M. and Schmidt, R., 1998. CBR in Medicine. In Case-Based Reasoning Technology (pp. 273-297). Springer, Berlin, Heidelberg.
10. Blanco, X., Rodríguez, S., Corchado, J.M. and Zato, C., 2013. Case-based reasoning applied to medical diagnosis and treatment. In distributed computing and artificial intelligence (pp. 137-146). Springer, Cham.
11. Vásquez-Morales, G.R., Martínez-Monterrubio, S.M., Moreno-Ger, P. and Recio-García, J.A., 2019. Explainable Prediction of Chronic Renal Disease in the Colombian Population Using Neural Networks and Case-Based Reasoning. IEEE Access, 7, pp.152900-152910.
12. Benamina, M., Atmani, B. and Benbelkacem, S., 2018. Diabetes diagnosis by case-based reasoning and fuzzy logic. IJIMAI, 5(3), pp.72-80.
13. COVID-19 WHO global report: https://covid19.who.int/ [Accessed 21 April, 2020]

14. Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K.S., Lau, E.H., Wong, J.Y. and Xing, X., 2020. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. New England Journal of Medicine.
15. Wu, J.T., Leung, K. and Leung, G.M., 2020. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. The Lancet, 395(10225), pp.689-697.
16. Johansson, M.A., Apfeldorf, K.M., Dobson, S., Devita, J., Buczak, A.L., Baugher, B., Moniz, L.J., Bagley, T., Babin, S.M., Guven, E. and Yamana, T.K., 2019. An open challenge to advance probabilistic forecasting for dengue epidemics. Proceedings of the National Academy of Sciences, 116(48), pp.24268-24274.
17. Tuite, A.R. and Fisman, D.N., 2020. Reporting, epidemic growth, and reproduction numbers for the 2019 novel coronavirus (2019-nCoV) epidemic. Annals of Internal Medicine.
18. Mizumoto, K., Kagaya, K., Zarebski, A. and Chowell, G., 2020. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. Eurosurveillance, 25(10), p.2000180.
19. Liu, Y., Gayle, A.A., Wilder-Smith, A. and Rocklöv, J., 2020. The reproductive number of COVID-19 is higher compared to SARS coronavirus. Journal of travel medicine.
20. Hu, Z., Ge, Q., Jin, L. and Xiong, M., 2020. Artificial intelligence forecasting of covid-19 in China. arXiv preprint arXiv:2002.07112.
21. Funk, S., Camacho, A., Kucharski, A.J., Eggo, R.M. and Edmunds, W.J., 2018. Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model. Epidemics, 22, pp.56-61.
22. Johansson, M.A., Apfeldorf, K.M., Dobson, S., Devita, J., Buczak, A.L., Baugher, B., Moniz, L.J., Bagley, T., Babin, S.M., Guven, E. and Yamana, T.K., 2019. An open challenge to advance probabilistic forecasting for dengue epidemics. Proceedings of the National Academy of Sciences, 116(48), pp.24268-24274.
23. Rao, A.S.R.S., Krantz, S.G., Kurien, T., Bhat, R. and Sudhakar, K., 2020. Model-Based Retrospective Estimates for COVID-19 or Coronavirus in India: Continued Efforts Required to Contain the Virus Spread. Current Science, 118(7), pp.1023-1025.
24. Buonomo, B., 2020. Effects of information-dependent vaccination behavior on coronavirus outbreak: insights from a SIRI model. Ricerche di Matematica, p.1.

25. Kim, S., Kim, Y.J., Peck, K.R. and Jung, E., 2020. School Opening Delay Effect on Transmission Dynamics of Coronavirus Disease 2019 in Korea: Based on Mathematical Modeling and Simulation Study. Journal of Korean Medical Science, 35(13).
26. Hellewell, J., Abbott, S., Gimma, A., Bosse, N.I., Jarvis, C.I., Russell, T.W., Munday, J.D., Kucharski, A.J., Edmunds, W.J., Sun, F. and Flasche, S., 2020. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. The Lancet Global Health.
27. Jia, J., Ding, J., Liu, S., Liao, G., Li, J., Duan, B., Wang, G. and Zhang, R., 2020. Modeling the control of COVID-19: Impact of policy interventions and meteorological factors. ELECTRONIC JOURNAL OF DIFFERENTIAL EQUATIONS Article Number: 23 Published: MAR 16 2020.
28. Choi, S. and Ki, M., 2020. Estimating the reproductive number and the outbreak size of Novel Coronavirus disease (COVID-19) using mathematical model in Republic of Korea. Epidemiology and Health, p.e2020011.
29. Huang, R., Liu, M. and Ding, Y., 2020. Spatial-temporal distribution of COVID-19 in China and its prediction: A data-driven modeling analysis. The Journal of Infection in Developing Countries, 14(03), pp.246-253.
30. Gostic, K., Gomez, A.C., Mummah, R.O., Kucharski, A.J. and Lloyd-Smith, J.O., 2020. Estimated effectiveness of symptom and risk screening to prevent the spread of COVID-19. Elife, 9, p.e55570.
31. Wang, C., Cheng, Z., Yue, X.G. and McAleer, M., 2020. Risk management of COVID-19 by universities in China. JOURNAL OF RISK AND FINANCIAL MANAGEMENT Volume: 13 Issue: 2 Article Number: 36 Published: FEB 2020.
32. Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J.M., Yan, P. and Chowell, G., 2020. Short-term forecasts of the COVID-19 epidemic in Guangdong and Zhejiang, China: February 13–23, 2020. Journal of Clinical Medicine, 9(2), p.596.
33. Yang, S., Cao, P., Du, P., Wu, Z., Zhuang, Z., Yang, L., Yu, X., Zhou, Q., Feng, X., Wang, X. and Li, W., 2020. Early estimation of the case fatality rate of COVID-19 in mainland China: a data-driven analysis. Annals of Translational Medicine, 8(4).
34. Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X., Huang, J., Shi, J., Dai, J., Cai, J., Zhang, T. and Wu, Z., 2020. Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. CMC-Computers, Materials & Continua, 63(1), pp.537-51.

35. Yin, F., Lv, J., Zhang, X., Xia, X. and Wu, J., 2020. COVID-19 information propagation dynamics in the Chinese Sina-microblog. Mathematical biosciences and engineering: MBE, 17(3), p.2676.
36. Zhou, W., Wang, A., Xia, F., Xiao, Y. and Tang, S., 2020. Effects of media reporting on mitigating spread of COVID-19 in the early phase of the outbreak. MATHEMATICAL BIOSCIENCES AND ENGINEERING Volume: 17 Issue: 3 Pages: 2693-2707 Published: 2020
37. Yang, C. and Wang, J., 2020. A mathematical model for the novel coronavirus epidemic in Wuhan, China. Mathematical Biosciences and Engineering, 17(3), pp.2708-2724.
38. Rong, X., Yang, L., Chu, H. and Fan, M., 2020. Effect of delay in diagnosis on transmission of COVID-19. Mathematical Biosciences and Engineering, 17(3), pp.2725-2740.
39. Hou, C., Chen, J., Zhou, Y., Hua, L., Yuan, J., He, S., Guo, Y., Zhang, S., Jia, Q., Zhao, C. and Zhang, J., 2020. The effectiveness of the quarantine of Wuhan city against the Corona Virus Disease 2019 (COVID-19): well-mixed SEIR model analysis. Journal of Medical Virology.
40. Prem, K., Liu, Y., Russell, T.W., Kucharski, A.J., Eggo, R.M., Davies, N., Flasche, S., Clifford, S., Pearson, C.A., Munday, J.D. and Abbott, S., 2020. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. The Lancet Public Health.
41. Ho, D., Addressing COVID-19 Drug Development with Artificial Intelligence. Advanced Intelligent Systems.
42. Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y. and Shen, D., 2020. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. In IEEE Reviews in Biomedical Engineering.
43. Kim, D., Hong, S., Choi, S. and Yoon, T., 2016. Analysis of transmission route of MERS coronavirus using decision tree and Apriori algorithm. In 2016 18th International Conference on Advanced Communication Technology (ICACT) (pp. 559-565). IEEE.
44. Dasgupta, S., Padia, A., Maheshwari, G., Trivedi, P., & Lehmann, J. (2018). Formal Ontology Learning from English IS-A Sentences. arXiv:1802.03701v1 [cs.AI].
45. Christopher, M. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 55-60.

46. Michelen, M., Jones, N. and Stavropoulou, C., 2020. In patients of COVID-19, what are the symptoms and clinical features of mild and moderate cases? Centre for Evidence-Based Medicine https://www. cebm. net/covid-19/in-patients-of-covid-19-what-are-the-symptoms-and-clinical-features-of-mild-and-moderatecase/accessed, 16.
47. Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C., & Napoli, R. D. (2020). Features, Evaluation and Treatment Coronavirus (COVID-19). StatPearls Publishing LLC.
48. Polamuri, S., 2015. Five most popular similarity measures implementation in python. Internet: http://dataaspirant. com/2015/04/11/five-most-popular-similarity-measuresimplementation-in-python.
49. Gan, M., Dou, X., & Jiang, R. (2013). From Ontology to Semantic Similarity: Calculation of Ontology-Based Semantic Similarity. Computational Systems Biology, 2013, 11.
50. Gu, D., Liang, C., & Zhao, H. (2017). A case-based reasoning system based on weighted heterogeneous value distance metric for breast cancer diagnosis. Artificial Intelligence in Medicine, 1-51.
51. Patel-Schneider, P.F. and Franconi, E., 2012, November. Ontology constraints in incomplete and complete data. In International Semantic Web Conference (pp. 444-459). Springer, Berlin, Heidelberg.
52. Herman, I., Horrocks, I., & Patel-Schneider, P. F. (2012, December 11). OWL 2 Web Ontology Language Document Overview (Second Edition). Retrieved April 19, 2020, from W3C: https://www.w3.org/TR/owl2-overview
53. O'Connor, M.J. and Das, A.K., 2009, October. SQWRL: a query language for OWL. In OWLED (Vol. 529, No. 2009).
54. O'Connor, M.J. and Das, A., 2009. SQWRL: a query language for OWL, OWL: experiences and directions (OWLED). In Fifth International Workshop (pp. 1-8).
55. Díaz-Agudo, B., González-Calero, P.A., Recio-García, J.A. and Sánchez-Ruiz-Granados, A.A., 2007. Building CBR systems with jCOLIBRI. Science of Computer Programming, 69(1-3), pp.68-75.
56. Kolodner, J.L., 1992. An introduction to case-based reasoning. Artificial intelligence review, 6(1), pp.3-34.
57. Althoff, K. (2012). Case-Based Reasoning and Expert Systems. Intelligent Information Systems Lab Institute of Computer Science University of Hildesheim.
58. Aamodt, A., & Plaza, E. (1994). Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications, 7(1), 39-59.

59. Ben-Bassat, M., Beniaminy, I., & Joseph, D. (1998). Combining Model-Based and Case-Based Expert Systems. Research Perspectives and Case Studies in System Test and Diagnosis Frontiers in Electronic Testing, 13(1), 179-205.
60. El-Sappagh, S., Elmogy, M., & Riad, A. M. (2015). A Fuzzy-Ontology Oriented Case Based Reasoning Framework for Semantic Diabetics Diagnosis. Artificial Intelligence in Medicine, 1-30.
61. Heras, S., Botti, V., & Juliana, V. (2013). A knowledge representation formalism for Case-Based-Reasoning. Agreement Technol, 105-119.
62. Rahim, R., Purba, W., Khairani, M. and Rosmawati, R., 2019, November. Online Expert System for Diagnosis Psychological Disorders Using Case-Based Reasoning Method. In Journal of Physics: Conference Series (Vol. 1381, No. 1, p. 012044). IOP Publishing.
63. Zhong, Z., Xu, T., Wang, F. and Tang, T., 2018. Text Case-Based Reasoning Framework for Fault Diagnosis and Predication by Cloud Computing. Mathematical Problems in Engineering, 2018.
64. Zhang, P., Essaid, A., Zanni-Merk, C. and Cavallucci, D., 2017. Case-based reasoning for knowledge capitalization in inventive design using latent semantic analysis. Procedia computer science, 112, pp.323-332.
65. Shen, Y., Colloc, J., Jacquet-Andrieu, A. and Lei, K., 2015. Emerging medical informatics with case-based reasoning for aiding clinical decision in multi-agent system. Journal of biomedical informatics, 56, pp.307-317.
66. Li, S.T. and Ho, H.F., 2009. Predicting financial activity with evolutionary fuzzy case-based reasoning. Expert Systems with Applications, 36(1), pp.411-422.
67. Petrovic, S., Mishra, N. and Sundar, S., 2011. A novel case based reasoning approach to radiotherapy planning. Expert Systems with Applications, 38(9), pp.10759-10769.
68. Fan, C.Y., Chang, P.C., Lin, J.J. and Hsieh, J.C., 2011. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Applied Soft Computing, 11(1), pp.632-644.
69. Begum, S., Ahmed, M.U., Funk, P., Xiong, N. and Von Schéele, B., 2009. A case-based decision support system for individual stress diagnosis using fuzzy similarity matching. Computational Intelligence, 25(3), pp.180-195.
70. Tversky, A., 1977. Features of similarity. Psychological review, 84(4), p.327.