Interestingness from COVID-19 Data: Ontology and Transformer-Based Methods

Abhilash C B[1], Nihar Sanda[1] and Kavi Mahesh[1]

[1]Indian Institute of Information Technology Dharwad (IIIT Dharwad), Karnataka, India

Abstract

Data Interestingness resides in most information systems. Significant implicit facts are hidden in healthcare data. Existing data interestingness techniques rely on standard data mining methodologies that lack the semantic aspect of the data. Data Interestingness is a useful functionality in analyzing large data corpora. Finding significant patterns in data helps make it more convenient and understandable for end users. In this study, our primary goal is to identify interesting patterns using ontology-based mining techniques and process them with BioClinicalBERT and CovidBERT to identify the interesting rules from the mined corpora. Further, we use the semantic similarity measure to compare the models with their similarity index to analyze the understanding of the model. The experimental results found that our proposed method is novel and operates on structured healthcare data using domain ontology. Finally, as a use case, we demonstrated using the proposed approach for paraphrasing the rules for decision-makers.

Keywords

Ontology, Semantic Annotation, Association Rule Mining, COVID-19, Interesting Patterns, Transformer Models

1. Introduction

The volume of healthcare data generated during the COVID-19 pandemic is having a significant impact on tabulating, summarizing, and indexing the facts that could help healthcare workers plan and prevent the spread, according to [1, 2]. Due to the accessibility of existing biomedical knowledge repositories, the current coronavirus pandemic highlights the need for automatic relation extraction techniques. In recent years, there has been a lot of time supporting the use of patterns in prediction models [3].

Machine learning models are rapidly becoming a powerful resource in healthcare. However, the quality of these models depends on the availability of high-quality training data. In addition to large datasets being necessary, these training sets must be robust and accurate. However, obtaining comprehensive and accurate real-world data for machine learning in healthcare is challenging due to privacy and ethical issues associated with such data.

Rule mining automatically mines the logical rules from a given knowledge base (KB). For example, the interesting rule mining methods find that “If A is the husband of B, and A lives in

WNLPe-Health 2022

abhilashcb@iiitdwd.ac.in (A. C. B); 19bcs125@iiitdwd.ac.in (N. Sanda); drkavimahesh@gmail.com (K. Mahesh)
https://github.com/abhilashcb8/ (A. C. B); https://koolgax99.github.io/ (N. Sanda);
https://iiitdwd.ac.in/Dr.Kavi_Mahesh.php (K. Mahesh)
0000-0001-9596-6630 (A. C. B); 0000-0001-5303-6400 (N. Sanda); 0000-0002-4353-725X (K. Mahesh)

the USA, then B also lives in the USA". This type of rule a mined based on certain confidence. These are necessary to have a complete KB. Rules are widely used in data and ontology for alignment and fact-checking purposes.

The Resource Description Framework (RDF) relies on graph-based structures. The description from a graph illustrates the relationships between the entities. Also, the information is decentralized, so connecting two graphs create a new graph. RDF follows an open-world assumption, facts that are stated are considered true, and the facts that are not stated are considered unknown

Motivation With the abundance of healthcare data available, it is critical for decision-makers to use it for predictive and preventive measures. Semantic data mining using ontology and transformer-based methods can reveal hidden inferences from data. This encourages decision-makers to keep track of data points that are relevant or interesting. The main focus of this paper is to infer interesting facts from two corpora of COVID-19 using the proposed interesting framework. More precisely, our contribution is as follows:

- We define a framework for data interestingness using domain ontology.
- We propose a novel technique to identify interesting rules using ontology and transformers-based methods.
- We compare the performances of two BioBERT Models for interestingness in COVID-19 data

Further, we demonstrated the usage of the proposed approach for paraphrasing the rules for decision-makers.

The remainder of the paper is as follows: Section 2 discusses the data and methods. Section 3 proposes the Ontology-based Data Interestingness (ODBI) framework used in this study. Section 4 discusses the results from two COVID-19 corpora by comparing their semantic nature of it. Section 5 concludes the paper by outlining future research directions.

2. Literature Study

Information extraction aims at automatically extracting information from unstructured data sources. Applications include information retrieval, opinion mining, sentiment analysis, question answering, and machine translation.

In computer science, the domain-specific task requires ontology as data and semantic model [1]. An ontology generally consists of an agreed (i.e., semantics) understanding of a specific field, axiomatization, explicitly expressed in a computer resource as a logical theory [4].

Association Rule Mining (ARM) is the most important topic in data mining research. Its goal is to discover interesting correlations, patterns, and associations between groups of items in transaction databases. Telecommunication networks, market and risk management, and inventory control all use association principles. Finding interesting association rules is a popular and current topic in data mining techniques [5, 6, 7].

In the state-of-the-art, several measurements are proposed, with ontology being less explored. An ontology that uses the semantic web, where data is represented as Resource Description Framework (RDF) triples (subject, predicate, object) makes it machine understandable. This fortifies the system to infer knowledge using the underlying schema of ontology [8]. The publication ”Attention is all you need” by [9] presented the Transformers architecture (2017). The architecture of transformers is encoder-decoder. The BERT model has recently produced cutting-edge results in a variety of NLP tasks in the same context. It’s a different kind of transfer learning. BERT’s primary operating mode is a transfer by fine-tuning similar to the one used by ULMFiT. Additionally, BERT can be used in the transfer mode by removing features like ELMo. Early detection model using Chat bot analytical language resources of descriptive questions to extract interesting facts. Three distinct models, CT-BERT, BERTweet, and Roberta are tuned on COVID-19-linked text data to distinguish between fake and real news [10].

Outcomes of Literature These successful studies demonstrate that ontologies can be used to improve the performance and enhance the usability of complex data analytics systems. The transformer models used in the study were pre-trained on biological data, giving them a deeper understanding of the terminology used in biomedicine. We use these state-of-the-art transformer-based methods for generating rule embeddings and cluster them further to analyze them with semantic scores for interesting ones.

3. Preliminaries, Data and Methods

This section explains the preliminary definitions and dataset with the proposed OBDI methodology.

3.1. Preliminaries

Ontology and ARM methods closely work towards the data interestingness [1]. In data mining literature, association rule mining is widely used for rule generation based on frequent patterns. This section aims to provide the readers with the necessary background knowledge.

Definition 1. *Association Rule: Technique used to mine the frequent patterns in Data. The discovered patterns define the relationship between them.*

we call **X→ Y** as association rule. To have the strong association rule, we need to compute the support and confidence as indicated in equations 1 and 2. Rules are defined considering our domain information.

$$Support = (X \rightarrow Y) = \frac{X \,\&\, Y}{Total\ number\ of\ attributes\ set} \tag{1}$$

$$Confidence(X \rightarrow Y) = \frac{Both\ X \,\&\, Y}{All\ value\ set\ containing\ X} \tag{2}$$

Definition 2. *Ontology: An Ontology **O** is defined as **O = (Tbox + Abox, G)**.
Tbox: define the schema or an ontology. Abox refers to RDF triples at the instance level. G is a labeled graph structure produced by connecting the relations with concepts. Figure 1 illustrates the importance of ontology.*

Definition 3. *Data Interestingness: Our notion of data is derived by integrating domain ontology with data in RDF and user interest rules.*

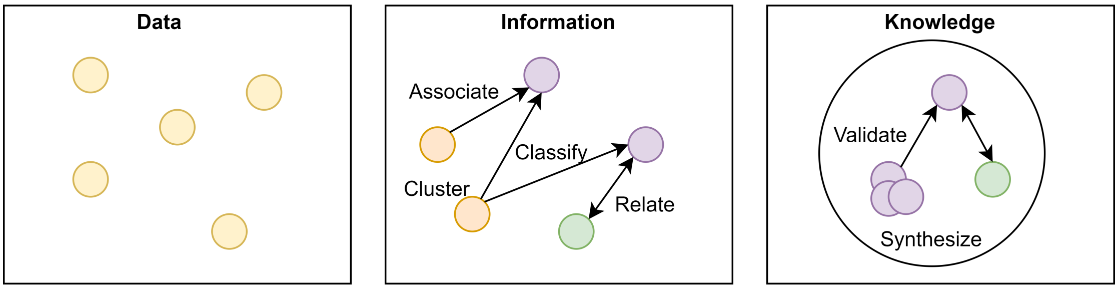


Figure 1: Ontology Illustration

Table 1
COVID-19 Dataset Descriptions

Dataset Name	**KATrace**
Description	The data is collected from HFWS web portal [3] and is curated and stored in spree,d-sheet by Siva Athreya and other researchers at the Indian Statistical Institute Bangalore. [4].
Attributes	Case ID, age, diagnosedOn, gender, city, cluster, reason, nationality, and status as attributes.
Data Download	www.isibang.ac.in/ athreya/incovid19/
Data Instances	71000
Dataset Name	**COP**
Description	The data is collected from the HFWS as part of the funded project. Data Access may be requested to HFWS.
Attributes	Case ID, age, Date, diagnosis, prescription for, drug store, district.
Data Download	Data Access may be requested to HFWS.
Data Instances	120000

3.2. Data

In this work, we used two COVID-19 corpora from the Indian state of Karnataka. [12] The data statistics is illustrated in Table 1.

3.3. Embeddings

We generate the embedding using the BioClinicalBERT and CovidBERT models in this study. BioClinicalBERT is a model trained with data corpora.

BioClinicalBERT is a model that is initialized on BioBERT (BioBERT-Base v1.0 + PubMed 200K + PMC 270K) and then it is trained on the MIMIC III [11] notes. These MIMIC notes consist of electronic health records from ICU patients of a hospital. For the pretraining of this

[1] https://karunadu.karnataka.gov.in/hfw/pages/home.aspx
[2] https://www.isibang.ac.in/ athreya/incovid19/

model, the authors utilized a batch size of 32, a maximum sequence length of 128 with a learning rate of $5 * 10^{-5}$. The models were trained for 150,000 steps using all MIMIC notes.

CoviBERT is a model that Deepset trains on AllenAI's COR19 dataset which consists of various scientific articles about coronaviruses. The model is initialized on BERT word piece vocabulary. Then, using the sentence-transformers library, it is fine-tuned on the SNLI and MultiNLI datasets to construct universal sentence embeddings using the average pooling technique and a softmax loss [12].

3.4. Methodology

The proposed OBDI framework, as in Figure 2, is an ontology-based mining framework that uses semantic similarity to determine the interestingness of rules. OBDI's goal is to automatically generate rules and knowledge from datasets to improve future decision-making process efficiency. OBDI's logic structure is as follows: RDF data instances are created from a dataset and a domain ontology. These data are backed by the domain experts' knowledge and also ontology concepts. Interesting rules are formulated as shown in Table 2,3 and 4 using the ontology and experts' knowledge. The OBDI methods include the IntApriori proposed [13]. It's significant that the generated rules are processed by BERT models for semantic scores to determine a rule's importance and degree of interest.

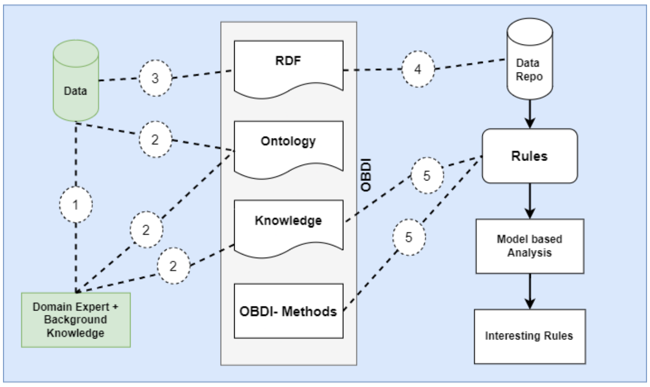
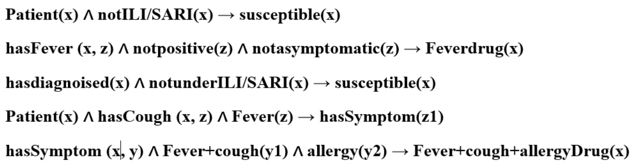


Figure 2: OBDI Methodology

4. Results and Discussions

This section discusses the semantic association rules generated from COVID-19 data and how it is processed on BioClinicalBERT and CovidBERT for identifying the similarity-based interesting rules. With the detailed analysis from the current state-of-the-art, BioClinicalBERT and CovidBERT are used in this study.

In a given set of rule embeddings, we find clusters that are closely related to the embeddings. This mapping is facilitated by BioBERT embeddings [14] of the rules generated by ontology-based mining. This helps reduce the rules' search space to have the most interesting ones. The



Patient(x) ∧ notILI/SARI(x) → susceptible(x)

hasFever (x, z) ∧ notpositive(z) ∧ notasymptomatic(z) → Feverdrug(x)

hasdiagnoised(x) ∧ notunderILI/SARI(x) → susceptible(x)

Patient(x) ∧ hasCough (x, z) ∧ Fever(z) → hasSymptom(z1)

hasSymptom (x, y) ∧ Fever+cough(y1) ∧ allergy(y2) → Fever+cough+allergyDrug(x)

Figure 3: Rules for COP ontology

cluster centroid is considered as the interesting point indicated as ***I***. The rules that match the cluster ***I*** value are termed as the most interesting ones. Focusing on the rules in the particular cluster, we use BioClinicalBERT and CovidBERT [15] embeddings and text summarization model to find the best-matched rules by generated the summarization of the cluster. Further, this summarization is treated as a paraphrase to decision-makers for future actions.

4.1. Semantic Association Rules from OBDI

The goal of the OBDI framework is to generate interesting rules, given the data and the domain ontology. The COP and COKPME ontology is used for generating semantic Association rules [13]. Our previous studies illustrate the design and implementation of COP and COKPME [5]. Table 2 shows the semantic association rules of the KATrace COVID-19 Dataset. Further, these rules will be used by BioClinicalBERT and CovidBERT to identify the interesting rules.

Rules in Ontology With the object and data properties defined in the COP ontology, the relationships inferred by the reasoner are the initial path for interesting fact generation. We define a set of rules that are operated on COP ontology for interesting fact generation. A few of the rules are indicated in Figure 3.

The rules are generated using ontology-based methods. A few of the rules with higher confidence are listed in Table 2. The generated rules are semantically annotated so that decision-makers can interpret them and take the appropriate actions. The results show the patient's age, status, the location from which he traveled, and the treatment provided.

Tables 3 and 4 describe the rules associated with its interesting index (I). The cluster centroid values are used as interesting data points, as are the embedding values that point to specific rules. Clustering using K-means [16] is applied to both the CovidBERT and BioClinicalBERT embedding sets. Both models generated five centroid points, five of which were interesting (I). Interesting rules are extracted from the rules pointing to the I value. The output of K-Means clustering on CovidBERT and BioClinicalBERT embeddings is shown in Figure 4.

Figure 5. depicts the distribution plot of the average of word embeddings obtained by the two models BioClinicalBERT and CovidBERT. The model's embedding distribution is also typical. The distributed rules demonstrate the model's understanding of the input rules. Table 5 also describes the ontology relationships distributed across the rule embeddings. It has been discovered that***treatmentProvided*** and ***sufferFrom*** are the two majorly identified ontology relationships.

[5] https://bioportal.bioontology.org/ontologies/COKPME

Table 2
Few Semantic Association Rules from COVID-19 KATrace Data Corpora

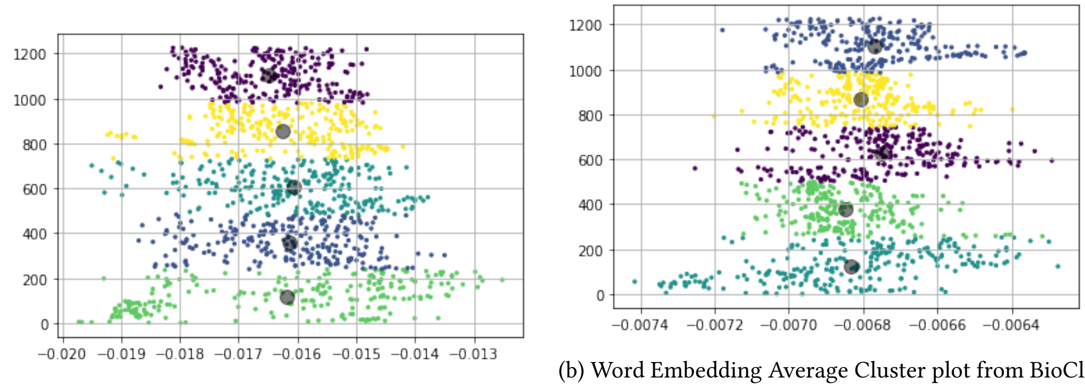
#	Semantic Association Rules of KATrace COVID-19 Dataset
R1	{sufferfromComorbidOthers}: (age, 27, 'Covid-19 (Suspect))⇒ (hasDiagnosedFor, Breathlessness(Influenza like Illness,)) (prescribedWith, Medicine Prescribed with Home Quarantine)
R2	{hasDiagnosedFor}: (age, above 65, Severe Acute Respiratory Infection) ⇒ (suspectedReasonOfCatchingCovid-19, Contact with other patients)
R3	{gender}: (Male, Female) (travelledFrom, TJ Congregation from 13th to 18th March in Delhi) ⇒ (suspectedReasonOfCatchingCovid-19, Family contact)
R4	{gender}: (Male, Female) ⇒ {(currentStatus ,cured) , (location, From Maharastra)
R5	{sufferFrom}(sneezing and an itchy, runny or blocked nose) ⇒ {prescribedWith}(Allergy Drugs)
R6	{sufferFrom}(sneezing and an itchy, runny or blocked nose, sore throat) ⇒ {sufferFrom}(Allergy Drugs, Cough Syrup)
R7	{sufferFrom}(Sweating, Headache, Muscle aches, Loss of appetite, Dehydration, General weakness, sore throat) ⇒ {sufferFrom}(Fever Drugs, Cough Syrup)

Table 3
Cluster Rule Summary with Interesting - I Value from CovidBERT

Cluster	Score (I)	Rule No.	Rule
C1	-0.0160542651	607	sufferfrom sneezing and an itchy, runny or blocked nose, sore throat treatmentprovided covid 19 testing
C2	-0.0161816583	121	hasDiagnosedFor, Breathlessness ILI prescribedWith Medicine Prescribed with Home Quarantine
C3	-0.0164832455	1101	hasage 29 sufferfrom fever treatmentprovided medicines prescribed with home quarantine advice
C4	-0.0162659895	853	hasage 0 hascategory ili sufferfrom fever treatmentprovided medicines prescribed with home quarantine advice
C5	-0.0161505655	364	hasage 29 sufferfrom fever hascategory ILI

Table 4
Cluster Rule Summary with Interesting-I Value from BioClinicalBERT

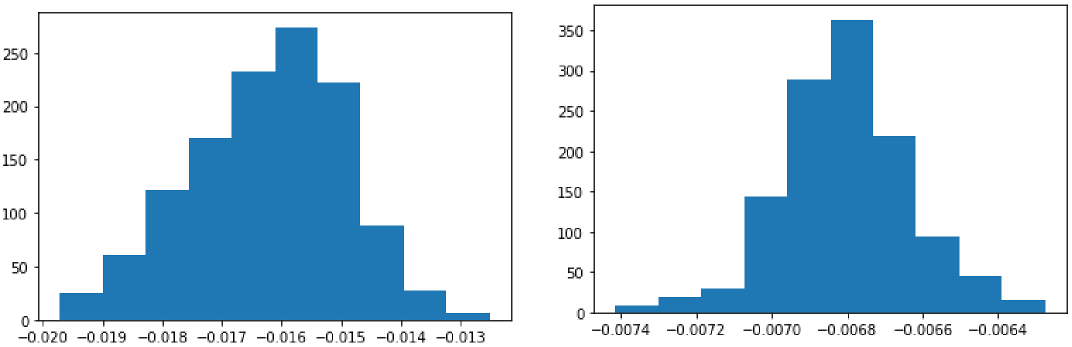
Cluster	Score (I)	Rule No.	Rule
C1	-0.00684050502	117	hascategory SARI treatmentprovided admitted to own hospital
C2	-0.00680626215	847	treatmentprovided referred to another hospital with call to emergency 108 or private ambulance
C3	-0.00676575871	598	livesin bangalore urban hascategory covid 19 suspect suggested for covid test
C4	-0.00677237246	1100	treatmentprovided call to emergency 108 for covid 19 testing hascategory covid 19 suspect
C5	-0.00683095187	355	hasage 27 treatmentprovided MP Home quarantine sufferfrom fever



(a) Word Embeddings Average Cluster plot from CovidBert

(b) Word Embedding Average Cluster plot from BioClinicalBERT

Figure 4: Word Embeddings Avg. Cluster plot from COVID and Clinical BERT Model



(a) Word Embedding Average plot from CovidBert

(b) Word Embedding Average plot from CLinicalBERT

Figure 5: Word Embeddings Avg. Distribution from COVID and Clinical BERT Model

The box-whisker plot of the semantic scores obtained from the two models BioClinicalBERT and CovidBERT is shown in 6. When compared to the BioClinicalBERT model, the CovidBERT model has a lot of variation in the semantic score. This demonstrates the two models' different levels of comprehension. The BioClinicalBERT model calculates high cosine similarity values between these rules, implying that they are very similar. The value of the min, max and mean similarity scores are as depicted in Table 6

The results show that the methodology learns to generate interesting facts based on the simple linguistic feature (COVID-19 Corpora) which are embedded in textual data using the BioClinicalBERT and CovidBERT model. The paraphrased summary of the identified interesting rules is as follows:

- Patient with {ILI, Diabetic} are highly prone to COVID-19 Infection.

Table 5
Relations Summary in each Rule

Relations	Count
prescribedwith	50
sufferfrom	812
hascategory	877
livesin	651
treatmentprovided	883
residesat	86

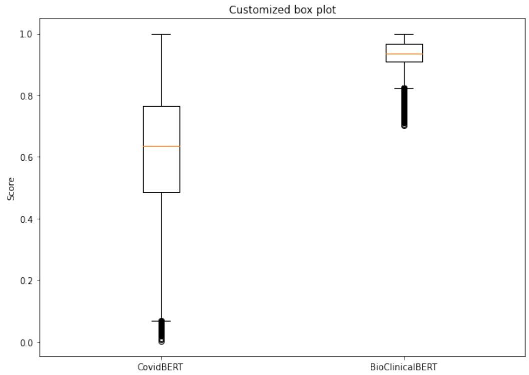


Figure 6: Box and Whisker Plot of Semantic Scores from BioClinicalBERT and CovidBERT.

Table 6
Semantic Similarity Scores for CovidBERT and BioClinicalBERT

Statistic	Semantic Score (BioClinicalBERT)	Semantic Score(CovidBERT)
Min	0.703100	0.003200
Max	0.999900	0.999300
Mean	0.933196	0.622179

- Below the age group of 35 is all suggested to have {MPHQ} advice. So healthcare facilities should be reserved for higher age groups.
- Many health workers are infected and admitted to their own hospitals, creating a shortage of resources.
- The most widely documented symptom in the COVID-19 dataset is the common flu.

The decision-makers understand these paraphrased rules for having preventive and predictive analysis.

5. Conclusion

This article proposes a novel methodology for mining ontology based on interesting facts from the COVID-19 data corpora. The Mined ontology-based rules are used as input to the transformer-based models like BioClinicalBERT and CovidBERT for interesting rules. The aggregate value of all rule embeddings is clustered. Next, using the cluster centroid, the Interestingness index (I) is derived and illustrated as the most interesting rule. Further, with the similarity scores from both models, the rules are compared for their similarity index. It observed that BioClinicalBERT outperformed CovidBERT with the similarity score by giving high relevance to the generated rules. As future research directions, this study is continued to compare the model-generated rules with domain expert rules to justify our claims. Another possible extension of this work could be to use it in the applications like the state-of-the-art COVID-19 Sentiment Analysis Toolkit.

6. Acknowledgments

This work was supported in part by the Department of Health and Family Welfare Services (HFWS), Government of Karnataka, India. We also extend our special thanks to the E-Health section of HFWS, Government of Karnataka, India, for providing all the necessary support and encouragement. Also, we would like to thank two anonymous reviewers for commenting on earlier versions of this paper.

References

[1] A. C, K. Mahesh, Graph analytics applied to covid19 karnataka state dataset, in: 2021 The 4th International Conference on Information Science and Systems, Association for Computing Machinery, New York, NY, USA, 2021, p. 74–80. URL: https://doi.org/10.1145/3459955.3460603. doi:10.1145/3459955.3460603.

[2] C. Abhilash, K. Mahesh, Ontology is what makes data interesting: Interestingness framework for covid-19 corpora, Journal of Information Science (2023).

[3] B. Bringmann, S. Nijssen, A. Zimmermann, Pattern-based classification: A unifying perspective, arXiv preprint arXiv:1111.6191 (2011).

[4] Wikipedia contributors, Fair data — Wikipedia, the free encyclopedia, 2021. URL: https://en.wikipedia.org/w/index.php?title=FAIR_data&oldid=1038845392, [Online; accessed 24-August-2021].

[5] F. H. AL-Zawaidah, Y. H. Jbara, A. Marwan, An improved algorithm for mining association rules in large databases, World of Computer science and information technology journal 1 (2011) 311–316.

[6] R. Agrawal, R. Srikant, et al., Fast algorithms for mining association rules, in: Proc. 20th int. conf. very large data bases, VLDB, volume 1215, Citeseer, 1994, pp. 487–499.

[7] F. F. dos Santos, M. A. Domingues, C. V. Sundermann, V. O. de Carvalho, M. F. Moura, S. O. Rezende, Latent association rule cluster based model to extract topics for classification and recommendation applications, Expert Systems with Applications 112 (2018) 34–60.

[8] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web, Scientific american 284 (2001) 34–43.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[10] A. Kumar, J. P. Singh, A. K. Singh, Covid-19 fake news detection using ensemble-based deep learning model, IT Professional 24 (2022) 32–37.

[11] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, Mimic-iii, a freely accessible critical care database, Scientific data 3 (2016) 1–9.

[12] S. Egami, Y. Yamamoto, I. Ohmukai, T. Okumura, Ciro: Covid-19 infection risk ontology, PloS one 18 (2023) e0282291.

[13] C. Abhilash, K. Mahesh, Ontology-based interestingness in covid-19 data, in: Research Conference on Metadata and Semantics Research, Springer, 2022, pp. 322–335.

[14] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

[15] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. McDermott, Publicly available clinical bert embeddings, arXiv preprint arXiv:1904.03323 (2019).

[16] C. Abhilash, K. Rohitaksha, S. Biradar, A comparative analysis of data sets using machine learning techniques, in: 2014 IEEE international advance computing conference (IACC), IEEE, 2014, pp. 24–29.