Ontology-Based Semantic Data Interestingness Using BERT Models

ARTICLE HISTORY
Compiled January 22, 2023

ABSTRACT
The study investigates the methods for finding hidden relationships and potentially useful facts in huge datasets. The COVID-19 pandemic has generated massive data in the healthcare sector in recent years, encouraging researchers and scientists to uncover the underlying facts. In line with this, our research proposes a novel framework for finding interesting facts from COVID-19 corpora using the proposed semantic interestingness framework. Since data mining with domain knowledge provides semantically rich facts, the proposed approaches use ontologies. The proposed framework uses the enhanced apriori algorithm for mining semantic association rules, and further, interesting rules are determined using BERT models for semantic richness. Further, to support our claims, we compared the outcomes of the proposed framework with the most recent approach in the field of data mining. As an evaluation mechanism for the rules, An evaluation framework is proposed that incorporates rule evaluation from domain experts and a Chi-Square test for statistical significance.

KEYWORDS
Semantic Mining, Semantic Interestingness, Association Rule Mining, Ontology-Methods, BERT Models.

1. Introduction

The COVID-19 pandemic is already a worldwide threat, demonstrating how susceptible humans may be. It has also inspired experts from various aspects and countries to find the potential solution to control the widespread. The volume of healthcare data generated during the COVID-19 pandemic significantly impacts tabulating, summarizing, and indexing the facts that could assist healthcare workers in planning and preventing the spread of the disease C and Mahesh (2021). The COVID-19 pandemic highlights the need for automatic relation extraction techniques due to the accessibility of existing biomedical knowledge repositories. In recent years the use of patterns in prediction models has been widely seen Bringmann, Nijssen, and Zimmermann (2011).

The pattern discovery method in data mining provides automatic detection of patterns. Patterns, in general, are the regularity, structure, or relationships in data, also known as association analysis Shawe-Taylor, Cristianini et al. (2004). Association Rule Mining (ARM) is the most crucial topic in data mining research. Its goal is to find interesting correlations, patterns, and associations between groups of items in transaction databases or other data repositories. 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 AL-Zawaidah, Jbara, and Marwan (2011). The *Apriori* algorithm family is based on two-rule extraction using support and confidence. Even though these two metrics are easy to compute, they generate a vast number of rules Srikant and Agrawal (1995), the majority of which are redundant and may be of no relevance to the user Agrawal, Srikant et al. (1994) dos Santos et al. (2018).

Furthermore, support and confidence only generate strong rules on their own. To extract interesting facts from data, additional measures along with support and confidence are required Manda, McCarthy, and Bridges (2013).

In the state-of-the-art, several measurements are proposed using ontologies in semantic mining. An ontology that uses the semantic web, where data is represented as Resource Description Framework (RDF), also referred to as triples (subject, predicate, object), makes it machine understandable. This fortifies the system to infer knowledge using the underlying schema of ontology Berners-Lee, Hendler, and Lassila (2001). In the semantic web, the information/data is organized in abstract form with its meaning. As a semantic data model, Ontology represents the COVID-19 data in entities (such as patient, diagnosis, treatment, locations) and relations between them sufferFrom, LivesAt, etc.). The Ontology for collection and analysis of COVID-19 data (CODO) ontology Dutta and DeBellis (2020) is used for the patient severity dataset, and the COVID-19 Ontology for Pharma (COP) ontology is used for COVID-19 COP Dataset Afolabi, Sowunmi, and Daramola (2017).

The Knowledge bases (KBs) created with the domain ontology can be mined for logical rules using Inductive Logic Programming (ILP), such as “If two people are identified in community spread of virus then, they (typically) live in the same city”. Large knowledge bases have been created due to recent improvements in information extraction. These knowledge bases include information, generally like “Delhi is India’s capital”, “Narendra Damodardas Modi was born in Vadnaga”, and “Every engineer is a person”. YAGO Suchanek, Kasneci, and Weikum (2007), DBpedia Bizer et al. (2009) are KBs that contain formation about a wide range of entities, including people, countries, cities, particular institutes, essential sites, and so on. The KBs know who was born where, who starred in which film, and who is the state’s chief minister. etc.

The general ARM techniques generate rules as follows: Considering the I = $\{I_1, I_2, ..., I_n\}$ be set of items then, group of items S = $\{S_1, S_2, ..., S_n\}$ such that S is a subset of I. The group set association rule A → B is defined over group G. The association rule is described with interesting metrics like support, confidence, and Lift. The different thresholds for these metrics will help to derive the interesting rules from the dataset Agrawal, Srikant et al. (1994). In our study, we use constraint-based pattern mining to determine the interestingness of data. The traditional Apriori algorithm is modified to prune the generated rules based on concise, reliable, and coverage measures. Additionally, we use transformer-based methods to identify the most interesting rules using the cosine similarity metric.

The publication “Attention is all you need” by Vaswani et al. (2017) presented the Transformers architecture (2017). The architecture of transformers is encoder-decoder. The Google AI team developed Bidirectional Encoder Representations from Transformers (BERT), a transformer-based pre-trained model Devlin et al. (2018). In this work, the semantic scores are employed as measures of importance, and their distributions, considering the distance measure, are calculated using the BERT models.

More precisely, our contributions are as follows:

- An effective data preprocessing technique that introduces semantics at the level of data curation.
- Semantic Interestingness Framework (SIF) for COVID-19 Data.
- Enhanced apriori approach with constraints (ConstApriori algorithm), which employs interestingness measures for semantic facts extracted from RDF data.

- Implementation of Clinical BERT and Bio BERT model for identifying the most interesting rules using a cosine similarity measure - Semantic Interestingness.

Also, the proposed evaluation framework validates the generated rules and justifies our claims.

The remainder of this paper is organized as follows: Section 2 summarises the existing methods and techniques. Section 3 describes the data, preliminaries, and data pre-processing techniques. Section 4 depicts the Semantic Interestingness Framework (SIF). Section 5 discusses the results and compares the semantic-rich rules to the state-of-the-art results. Section 6 discusses the implementation of the BERT models for semantic interestingness. In Section 7, a significance test and domain expert evaluation are performed. Section 8 concludes with future research directions.

2. Related Work

In computer science, the domain-specific task requires ontology as data and semantic model C and Mahesh (2021). 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 Wikipedia contributors (2021).

COVID-19 ontology for cases and patient information (CODO) Dutta and DeBellis (2020) is a model designed to collect and analyze COVID-19 data. The ontology is standard-based and can incorporate data from multiple sources. New ways for selecting meaningful association rules based on a variety of metrics are defined in the literature Badenes-Olmedo et al. (2020) He et al. (2020) Agrawal, Srikant et al. (1994).

Prior research on COVID-19 has concentrated on forecasting case numbers Arora, Kumar, and Panigrahi (2020); Qin et al. (2020); Tomar and Gupta (2020) and categorising COVID-19 patients from real-world x-ray data sets using sophisticated deep neural network techniques Apostolopoulos and Mpesiana (2020); Ozturk et al. (2020). These techniques, however, focus on examining COVID-19 symptom patterns.

In recent literature on interestingness measures for ontology methods, various domains are considered, like research on Interestingness Measures (IM's), Semantic data mining, Ontology matching, Hierarchical measures, and different rule pruning techniques. Several algorithms have been designed considering threshold as a measure Manda, McCarthy, and Bridges (2013) Geng and Hamilton (2006).

In Bellandi et al. (2007), Bellandi et al. demonstrated how ontologies could improve the rules obtained by ARM systems. The post-processing of ARM results using an ontology for consistency testing is presented by Marinica and Guillet (2010). Filtering the identified rules is proposed by Mangla and Akhare (2015). This method takes advantage of the user's and domain expert's knowledge. It combines user knowledge with ontologies linked to data in post-processing. Moreno, Segrera, and López (2005) suggested an algorithm that reduces the number of pruning operations in the Apriori algorithm. They used *apriori-gen* operation to produce the candidate 2-item sets.

Ignoring the coded knowledge at the schema level has an adverse impact on the interpretation of the discovered rules. Barati, Bai, and Liu (2016) propose SWARM (Semantic Web Association Rule Mining) that automatically mines Semantic Association Rules. Shan, Zhou, and Zhang (2021) and Biradar, Saumya, and Chauhan (2022) used the transformed-based models for fake news detection. It's observed that the model specific to the healthcare domain yields better results compared to generalised ones. In Alzubi et al. (2021) COBERT-question answering system design was created to

Table 1. BERT Model for Healthcare Domain

Model	Trained Corpora	Learning Type	Hyperparameters	NLP Task
ClinicalBERT	MIMIC-III dataset	Supervised/unsupervised	Sentence length: 128–512 tokens	1,2,3
BlueBERT	PubMed abstracts and MIMIC-III clinical notes	Supervised/unsupervised	Sentence length: 128–512 tokens	1,2,3
CovidBERT	COVID Tweets	Supervised/unsupervised	Sentence length: 128–512 tokens	1,2,3
BioClinicalBERT	large-scale biomedical corpora (PubMed)	Supervised/unsupervised	Sentence length: 128–512 tokens	1,2,3
CovBERT	Cov-Dat-20 (PubMed, Scientific articles)	Supervised/unsupervised	Sentence length: 128–512 tokens	4

1. Biomedical Named Entity Recognition, 2. Biomedical Relation Extraction, 3. Biomedical Question Answering, 4. Text Classification.

address COVID-19 difficulties and quickly assist researchers and clinical professionals in obtaining legitimate scientific information. For COVID-19-related question answering, cosine similarity measures are applied on the word embedding for categorizing the top-K documents Choi et al. (2018) Shen et al. (2020) Guo et al. (2020).

The detailed literature review found that the ontology-based approach is extensively used for explicit and implicit fact generation. However, to the best of our knowledge, ontology-based approaches for rule generation have relatively little attention. While mining semantic association rules using ARM is proposed. Even if threshold-based rule pruning is specified, it is uncertain whether the remaining rules may be combined due to their semantic nature for finding interesting rules. To overcome these shortcomings, our proposed model uses the richness of ontology methods by using schema and RDF instances with ARM techniques to find interestingness in data. Because of transformer-based methods, it's generally used in question-answering and sentiment analysis methods. Our methodology adopts it to have the most interesting rules using cosine similarity-based measures.

3. Preliminaries, Data, and Pre-processing Technique

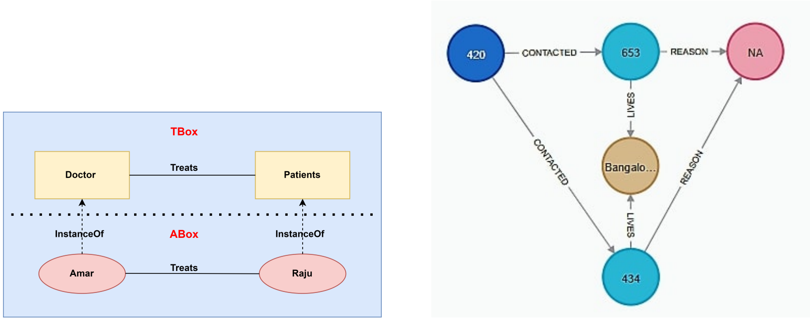
This section discusses the dataset with preliminaries and an effective data processing technique designed for this study. Besides the general data analysis and knowledge engineering methods, the ontology-based approach stands aside and has unique importance.

3.1. Preliminaries

Ontology and ARM methods closely work towards the data interestingness C and Mahesh (2021). In data mining literature, association rule mining is widely used for rule generation based on frequent patterns.

Definition 3.1. Association Rule: Technique used to mine the frequent patterns in Data. The discovered patterns define the relationship between them.

we call $\mathbf{X} \rightarrow \mathbf{Y}$ 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. AVS refers to the attribute value set.



(a) Example of ABox and TBox of an Ontology.

(b) Knowledge graph view from COVID-19 Data

Figure 1. Representation of ontology and knowledge graph.

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

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

Definition 3.2. 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 labelled graph structure produced by connecting the relations with concepts.

Figure 1 represents the ontology and a COVID-19 knowledge graph snippet.

Definition 3.3. Knowledge Graph: A collection of descriptions of concepts, things, relationships, and events that are all linked together.

Definition 3.4 (Data Interestingness (DI))**.** Our notion of Data Interestingness is derived by integrating domain ontology *(O)* with data in RDF *(D)* and with user interest rules *(U)* as shown in equation 3.

$$\boldsymbol{DI} = \{\boldsymbol{O,D,U}\} \tag{3}$$

Here user interest(U) refers to the unexpectedness and actionabILIty measures.

Definition 3.5. Semantic Interestingness (SI): Centroids of the semantic score cluster

3.2. COVID-19 Corpora

In this work, two COVID-19 corpora from the Indian state of Karnataka are used. . The interesting framework is introduced based on domain ontology to get interesting facts. The KATrace (COVID-19) dataset is open-source data, and the COKPME (COVID-19) dataset is from the Ministry of Health and Family Welfare Service (HFWS), Karnataka Government, for research purposes only. The data is structured with patient

[0] https://karunadu.karnataka.gov.in/hfw/pages/home.aspx

Table 2. COVID-19 Corpora Description.

Dataset	Description	Features
KAtrace	71000	8
COKPME	250000	6

[a] Common features across this dataset are Symptoms and Diagnosis.

Table 3. COVID-19 Dataset Descriptions.

Dataset Name	**KATrace**
Description	The data is collected from HFWS web portal [2] and is curated and stored in spree,d-sheet by Siva Athreya and other researchers at the Indian Statistical Institute New York. [3].
Attributes	Case ID, age, diagnosedOn, gender, city, cluster, reason, nationality, and status as attributes.
Data Download	www.isibang.ac.in/ athreya/incovid19/
Dataset Name	**COKPME**
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.

demographic and clinical symptom details. The data from HFWS was provided by annomyzing the patient demographic details. [1] The data statistics are illustrated in Table 2 and 3.

3.3. Data Pre-processing Technique

The most crucial step in the interestingness framework is data pre-processing, referred to as RDF data processing. The structured data is translated to RDF triple form and uses the SPARQL Endpoint for query operations. The steps of the proposed data pre-processing technique are as follows:

- Data Curating
- Converting to RDF
- Linking Data Sources
- Publishing as Knowledge Graph

3.3.1. Data Curation

Introducing semantics is a unique technique in data curation. Semantics acts as a glue by defining the data model and relationships before and after the data processing. Curation handles data duplication and improves data quality. It disambiguates the data items. Multiple labels are unified and classified based on the trained model corpora, and semantically equivalent terms are identified.
Example: {Fever, Fiver, FVR, Feever, Fevere} = **Fever**

For the data curation, first, we create the corpus of the words from the COVID-19 data illustrated in Table 3. Now, the data corpus has many misspelled English words and clinical abbreviations. To handle the misspelled English words, The ***SymSpell's*** own corpus is used, which consists of the intersection of the large Ngram datasets from

[1] https://karunadu.karnataka.gov.in/hfw/pages/home.aspx

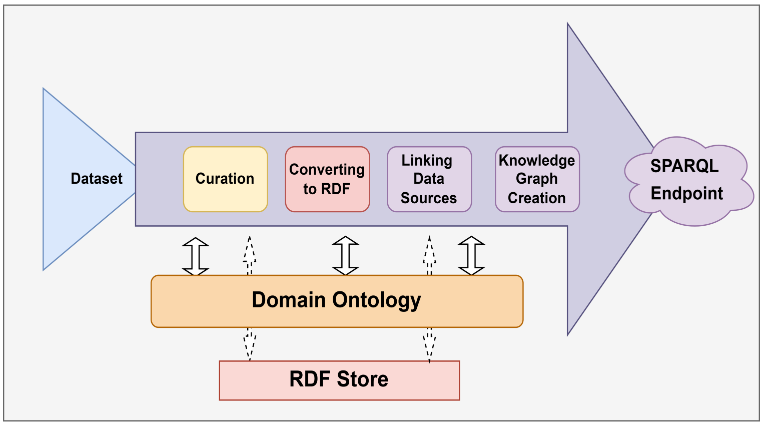


Figure 2. Data Pre-processing Technique

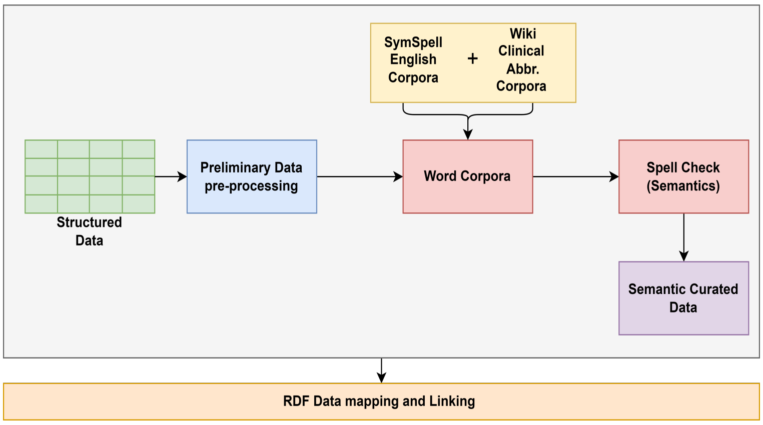


Figure 3. Semantic Data Curation

google Ngram with wordlists generated from ***hunspell*** dictionary files. The proposed semantic data curation is pipelined as shown in Figure 3. For the clinical abbreviations, we generated the word corpus of clinical abbreviations. The corpus was web scrapped from Wikipedia's Page on List of Medical Abbreviations. To get the curated semantic data of the original corpus, we then run a Garbe (2012) ***Symspell*** check algorithm using the intersections of both the corpus and subsequently achieving the semantic data curation at the initial level of RDF data processing. [4]

3.3.2. Converting to RDF

Datasets are converted to uniform representation, RDF. As discussed earlier, RDF makes a structured data representation with its relationships. The data pipeline program is used in Figure 2 to convert the curated data into RDF format.

3.3.3. Linking Data Sources

Linking data is the best practice, and it is done using the web links information using RDF and IRIs. The data linking is done using domain ontology. The ontology concepts

[4] https://en.wikipedia.org/wiki/Listofmedicalabbreviations

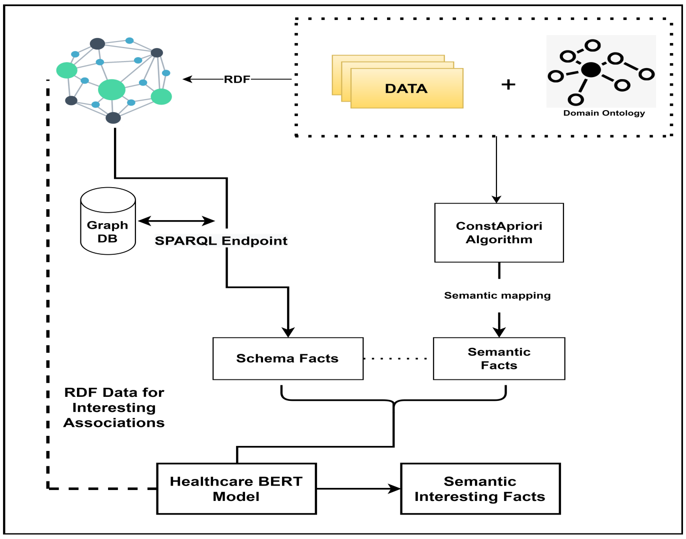


Figure 4. Semantic Interestingness Framework

with the prefix and URI are used to link the data sources. In our study, we have used [5] , [6] , [7] , [8] , [9] data sources.

3.3.4. Publishing as Knowledge Graph

The knowledge graph is installed on a Windows PC using the GraphDB tool and the SPARQL Endpoint for query processing. SPARQL, a query language similar to SQL, is used to query the knowledge graph (Triple store). Overall, it constitutes a COVID-19 knowledge base or a COVID-19 triple store.

4. Interestingness Framework

The proposed framework in this study has two directions for data interest: schema level and instance level. To extract scheme-level interestingness, COP ontology relationships are utilized. The intrinsic attributes are extracted and stored as S-Rules by the schema-level facts. For example, we employ RDF data instances generated using the domain ontology. Furthermore, the semantic annotation for the created rules is performed using domain information from the ontology. The interestingness framework is used to build semantically rich rules. Figure 4 depicts the methodology used by the semantic association rule to derive interestingness in data.

4.1. Motivating Example

The motivation behind this work is explained in this subsection.

- John lives in New York.

[5] http://www.w3.org/1999/02/22-rdf-syntax-ns
[6] http://www.w3.org/2002/07/owl
[7] http://iiitdwd.ac.in/cop
[8] http://www.w3.org/2000/01/rdf-schema
[9] http://xmlns.com/foaf/0.1/

Algorithm 1 Algorithm for ConstApriori (CBA)

```
Input: Rule Set (RS), Constraint (C);
Output: Set of Interesting rules
  Initialize : min_sup, min_conf
a = Concise(RS)
b = Reliable(RS)
c = Coverage(RS)
RP = a+b+c
for each rule R in the RP do
  if R $\geq$ min_sup AND R $\geq$ min_conf
    TR $\leftarrow$ R
    continue
  end if
end for
for each constraint C in TR
  Look for constraint C in input TR
   if found
     Store rules in R to the TR
     Continue;
   else
     go on with next R
end for
```

- John suffers from Diabetic and Hypertension.
- Jack had a cardiac problem for the past year.
- John and Jack are treated at Apollo Healthcare centre.

Now, on the above facts, inferences are as follows:

- Jack was treated in New York.
- Apollo Healthcare centre is located in New York.
- Diabetic, Hypertension, and Cardiac-related treatments are provided at Apollo Healthcare centre.

The association rule mining algorithm is primarily concerned with detecting frequent patterns. The classic apriori algorithm is employed based on association rule mining. The enhancement is done using the constraint-based method - *ConstApriori* algorithm 1, for mining constraint-based patterns. These patterns are better matched to the preferences of the user. Using the domain ontology, the resulting constraint-based rules are semantically annotated. The semantic annotation is done concerning the ideas and relationships in the ontology.

4.2. Interesting Measures

Definition 4.1. Concise item: A Concise Item C_i is a pattern, i.e., C_i = (S, O). S is the subject element, and O is the object element. An item is concise if it contains relatively few attribute pairs. It contains a list of subjects or objects, i.e., $\{S_1, S_2, ..., S_n\}$ or $\{O_1, O_2, ..., O_n\}$ with minimum pair combinations. Corresponding S, O contains a combination of predicate object or predicate-subject, i.e., (P, O) or (P, S).

Example

{residesAt: New York_Urban → sufferFrom(ILI)} {hasAge: 45, sufferFrom(COVID-19 Illness) → otherComorbidity(Diabetic)}

Procedure 1 Concise Item

```
Input: S, (where S=dataset of transactions), min_sup = val, min_conf =val,
    min_lift=val, min_len=val;
Output: Set of frequent patterns with the initialized min_len
    Initialize : FI=Ø
for all pattern set p in S do
  if support $\geq$ min_sup then
    for each pattern set p in S do
      count $\leftarrow$ read.count()
      if count $\leq$ min_len then
        FI $\leftarrow$ p
      end if
    end for
  end if
end for
```

Definition 4.2. Reliable item: A reliable item set R_i contains a set of reliable items with a consistent set of items. The entity sets E_S, i.e., $R_i = \{S_1, S_2, ..., S_n\} = (SemE_s + SemP_s)$, where E_s is entity set and P_s is property set. Items can be aggregated to make them reliable items.

Procedure 2 Reliable Item

```
Input: S, (where S=dataset of transactions), min_sup = val, min_conf =val,
    min_lift=val, min_len=val;
Output: Set of frequent patterns with the higher count value
    Initialize : FI=Ø
for all pattern set p in S do
  if support $\geq$ min_sup then
    for each pattern set p in S do
      count $\leftarrow$ read.count()
      if count $\geq$ min_conf then
        FI $\leftarrow$ p
      end if
    end for
  end if
end for
```

Definition 4.3. Coverage item: A coverage item is considered covered in general if it applies to a reasonably large subset of a dataset. It contains a set of semantic items with similar element sets. i.e., CO $= \{S_1, S_2, ...S_n\} = \{E_s, P_s\}$ where E_s = Entity set and P_s is Property set. $CO_i = \{E_s \bigcup P_s\}$. Items can be aggregated to infer new knowledge that covers large amounts of data.

Example

{residesAt: New York_Urban → sufferFrom(ILI)} {Sup:0.60 Conf:0.75}
{residesAt: New York_Urban, sufferFrom(COVID-19 Illness) → otherComorbidity(Breathlessness)

Procedure 3 Coverage Item

```
Input: S, (where S=dataset of transactions), min_sup = val, min_conf =val,
    min_lift=val, min_len=val;
Output: Set of frequent patterns with the initialized $min_len$;
    Initialize : FI=Ø, len=Ø
for all pattern set p in S do
  if support ≥ min_sup then
    for each pattern set p in S do
      if len ≥ min_len
        FI ← p
      end if
    end for
  end if
end for
```

5. Result and Discussions

The experiment result was executed on Intel(R) Core(TM) i7-8565U CPU @ 1.80GHz machine with 16GB RAM. The software platforms and tools used include Protégé for designing the owl ontology and generating RDF data and the Neo4j- graph database tool for processing the Cipher query on RDF data. The rule generation and interestingness mining algorithms discussed in section 4 are coded in python on google collaboratory. Ontology documentation by WIDOCO tool, visualization by WebVOWL. Garijo (2017) Musen (2015). [10] [11]. The proposed interestingness framework measures the quality of discovered rules by taking information at both the instance and schema levels into account. It generates semantically-enriched rules by using (rdf:type, rdfs: subClassOf).

Table 4 and 5 indicate the two COVID-19 corpora used to evaluate the proposed framework for interestingness using the three pruning techniques. Figures 5 depict the rule distribution regarding the frequency of occurrence. It is observed that in both data corpora, the distribution is normal. [12] [13].

Table 6 refers to a constraint file example that is used to mine the user-preferred

[10] http://protege.stanford.edu

[11] https://www.dbpedia.org/

[12] https://covid19.karnataka.gov.in/english

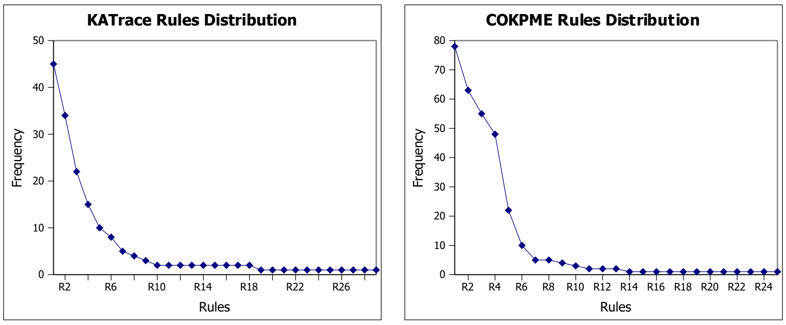
[13] https://www.mohfw.gov.in/

Table 4. KATrace Rules Summary

#	MinSup	MinConf	MinLift	MinLen	MaxLen	No. of Rules
Traditional Method	0.50	0.60	2	2	4	1153
Concise	0.50	0.60	2	2	2	286
Reliable	0.50	0.60	2	2	3	123
Coverage	0.50	0.60	2	2	3	90
Traditional Method	0.40	0.50	2	2	4	1275
Concise	0.40	0.50	2	2	2	374
Reliable	0.40	0.50	2	2	3	145
Coverage	0.40	0.50	2	2	3	110
Traditional Method	0.60	0.80	2	2	4	457
Concise	0.60	0.80	2	2	2	96
Reliable	0.60	0.80	2	2	3	56
Coverage	0.60	0.80	2	2	3	45

Table 5. COVID-19 COKPME Rules Summary

#	MinSup	MinConf	MinLift	MinLen	MaxLen	No. of Rules
Traditional Method	0.50	0.60	2	2	3	1784
Concise	0.50	0.60	2	2	3	298
Reliable	0.50	0.60	2	2	3	541
Coverage	0.50	0.60	2	2	3	389
Traditional Method	0.60	0.80	2	2	3	790
Concise	0.60	0.80	2	2	3	178
Reliable	0.60	0.80	2	2	3	127
Coverage	0.60	0.80	2	2	3	248



(a) Rules frequency of top 25 rules of KATrace Dataset.

(b) Rules frequency of top 25 rules of COKPME Dataset.

Figure 5. Rules frequency of top 25 rules.

rules. The *ConstApriori* algorithm filters the rules based on the semantic key terms illustrated in the constraint file using the constraint file as input. To begin, *ConstApriori* mines frequent patterns from RDF data and then uses the constraint file to generate interesting rules.

5.1. 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 operating on COP ontology for interesting fact generation. A few rules are indicated in Figure 6.

To generate Semantic rules, we use the SPARQL query. The RDF triple store is hosted on GraphDB with http://localhost:7200/ as SPARQL Endpoint. Few SPARQL queries are illustrated below:

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX cokpme: <http://www.iiitdwd.ac.in/ns/cokpme#>
PREFIX schema: <https://schema.org/>
SELECT ?predicate (COUNT(*)AS ?frequency)
WHERE {?subject ?predicate ?object}
GROUP BY ?predicate
ORDER BY DESC(?frequency)
LIMIT 10
```

```
PREFIX rank:<http://www.ontotext.com/owlim/RDFRank#>
SELECT ?n
```

Table 6. Example of Constraint File

Constraint Key Terms
{Hypertension}, {mild and very mild covid-19}, {severe covid-19},{ moderate covid-19}, {Fever, allergy, cough}, {ILI}, {SARI}, {body ache}, {Depression}, {respiratory infection}, {Sore throat},{Vomiting} , {Abdominal pain},{Diarrhea}, {Difficulty in Breathing}, {Heart Disease}

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 6. Rules for COP ontology

```
WHERE {?n rank:hasRDFRank ?r }
ORDER BY DESC(?r)
LIMIT 100
```

The most atomic level of information that can be in a graph is a binary interaction, whereas, in RDF, the binary interaction is decomposed into triplets. For instance, identifying all those with a close relative diagnosed with COVID but not yet tested is indicated in Table 8.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX codo: <http://www.isibang.ac.in/ns/codo#>
PREFIX schema: <https://schema.org/>
SELECT ?p ?r
WHERE{
    ?p rdf:type schema:Patient.
    ?p codo:hasDiagnosis ?d.
    ?d rdf:type codo:COVID-19Diagnosis.
    ?p codo:hasCloseRelationship ?r.
    ?r codo:hasCovidTest false.
}
```

Table 7. SPARQL Output

Predicate	Frequency
:hasdiagnosedFor	”531978”^^xsd:integer
:medicineFor	”514014”^^xsd:integer
http://iiitdwd.ac.in/patient	”422610”^^xsd:integer
:address	”422574”^^xsd:integer
cokpme:address	”175003”^^xsd:integer
cokpme:addressLocality	”175003”^^xsd:integer
cokpme:diagnosedFor	”175003”^^xsd:integer
cokpme:foaf:age	”175003”^^xsd:integer
cokpme:sufferFrom	”175003”^^xsd:integer
cokpme:treatmentProvided	”175003”^^xsd:integer

Table 8. SPARQL Output

?p	?r
codo:PX000001	codo:PX000004
codo:PX000001	codo:PX000005
codo:PX000001	codo:PX000006
codo:PX000001	codo:PX000007
codo:PX000002	codo:PX000008
codo:PX000003	codo:PX000010
codo:PX000003	codo:PX000012

5.2. Semantic Interestingness in COVID-19 Corpora

The interestingness framework aims to generate interesting rules given the data and the domain ontology. The semantic association rules are extracted using the *ConstApriori* algorithm, which is based on the interestingness framework. The overall framework has two components: schema-based fact generation using SPARQL queries and RDF data on the proposed rule mining algorithm.

Table 9. Top 4 Semantic Association Rules of COVID-19 KATrace Dataset

#	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)

Table 10. Top 10 Semantic Association Rules of COKPME COVID-19 Dataset

#	Semantic Association Rules of COKPME Dataset
R1	{patient}: (sufferFrom, Diabetic) ⇒ (Influenza like Illness, Medicine Prescribed with Home Quarantine)
R2	{has Age, 35}: (sufferFrom, Influenza like Illness), (diagnosedFor, Breathlessness(Influenza like Illness)) ⇒ (prescribedWith, Medicine Prescribed with Home Quarantine)
R3	{Influenza like Illness}: (diagnosedFor, COVID-19 (Suspect)) ⇒ (Medicine Prescribed with Home Quarantine, Admitted to own Hospital)
R4	{breathlessness}: (diagnosedFor, SARI) ⇒ (Admitted to other Hospital, Call to Emergency for COVID-19 test)
R5	{'Covid-19 (Suspect)'}: (isTakenFor, 'Call to Emergency 108 for Covid-19 testing') ⇒ (sufferfromComorbid, breathlessness), (livesIn, Bangalore Urban))
R6	{'No Comorbid Conditions, skin rashes, 'Anxity'}) ⇒ (NOTHING SIGNIFICANT)
R7	{hasAge, 27}: (sufferfromComorbidOthers, 'Covid-19 (Suspect')), (Breathlessness, diagnosedFor(Influenza like Illness,)) ⇒ (prescribedWith, Medicine Prescribed with Home Quarantine)

The traditional method rules are derived from work on the COVID-19 dataset using association rule mining Tandan et al. (2021). The rules appear interesting because the work covers a variety of attributes such as age, gender, and symptom for generating patterns. However, the proposed method mines interesting patterns using constraint-based and SPARQL query-based methods, making it more interesting. In addition, the rules are annotated with ontology concepts and relationships. This makes the proposed method rules more interesting and understandable to decision-makers.

Table 11. Comparative Analysis of Rules

Traditional Method Rules	Proposed Method Rules
{Breathing problem, Sputum} ⇒ {Cough}	{treatmentprovided(Admitted to own hospital)} ⇒ {sufferFrom(Fever)}
{Respiratory failure, Septic shock} ⇒ {Pneumonia}	{hasCategory(ILI) ⇒ {treatmentprovided(Admitted to own hospital)}
{Cardiac arrythmia, Renal disease} ⇒ {Respiratory distress syndrome}	{hasAge(18)}, {hasCategory(ILI)} ⇒ {sufferFrom(Fever)}
{Breathing problem, Respiratory distress syndrome} {Died}	{hasAge(0)}, {sufferFrom (Fever)} ⇒ {hasCategory(ILI)}
{Fever, Heart failure} ⇒ {Cough}	{hasAge(55)} ⇒ {hasCategory(ILI), {treatmentprovided(Medicines prescribed with home quarantine advice), {sufferFrom(Fever)}
{Cardiac arrythmia, Septic shock} ⇒ {Died}	{hasCategory(SARI), {treatmentprovided(Referred to another hospital with call to emergency 108 or private ambulance)} ⇒ {sufferFrom(Fever, Cough)}
{Headache, Malaise/body soreness, Weakness} ⇒ {Male}	{livesIn(New York urban) ⇒ {treatmentprovided(Admitted to own hospital)}, {hasCategory(ILI)}

6. Semantic Interestingness using Transformers

To choose the appropriate BERT model for the semantic interestingness technique, we did a thorough literature study as indicated in Table 1. We decided to leverage the pre-trained transfer learning models as they have significant relevance to the domain and the trained corpora. According to the study, we conclude that Transformer-based methods are much less explored to their full potential to detect interestingness in the rules. Two such models, BioClinicalBERT and CovidBERT, are selected for this research work to find the most interesting rules from the input rule set. We must design a practical and reasonable model that justifies our claim for interestingness. With GPU-enabled computing resources, these pre-trained models are fine-tuned using the Simple Transformers library on Google Collaborator. The detailed workflow of the transformer-based interestingness method in our study is shown in Figure 7. The rules obtained from semantic association rule mining are processed using BioClinicalBERT and CovidBERT models to generate rule embeddings. Next, using cosine similarity measures for a total input of 1242 unique rules, we generated 752151 rules with a similarity index. This huge corpus is obtained after mapping each rule with a different rule. Further, we identify the interesting centroids by applying clustering on the average rule embeddings and cosine similarity index of all the rules. The centroid is computed in two aspects of interest, first on the average value of the rule embeddings score and second on the cosine similarity index of all the rules. Both aspects are compared for interestingness in data.

6.1. Tokenization and Embeddings

All of the pre-trained models require a specific format for the input text. In our case, for the semantic association rules, tokenizers break the input text into smaller tokens. Before fine-tuning the model, these tokens are then transformed into embeddings. Each transfer learning model has undergone pre-training on a particular corpus with a predetermined vocabulary. The model's input text may include terms not part of its set vocabulary. The normal BERT model employs a WordPiece tokenizer to handle these terms that are out of the vocabulary (OOV) while retaining information from the input data Schuster and Nakajima (2012). It is trained on lower-cased English text

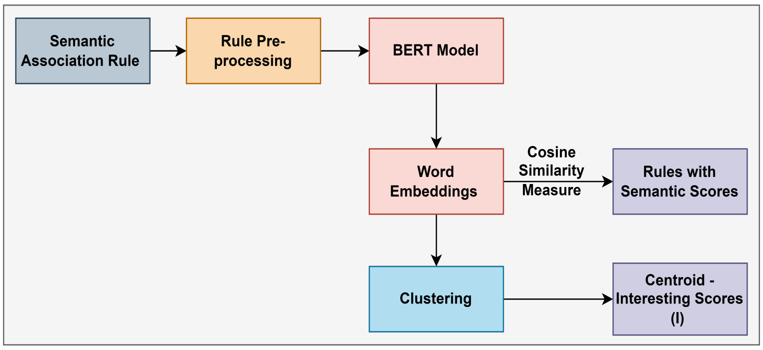

Semantic Association Rule
Rule Pre-processing
BERT Model
Word Embeddings
Cosine Similarity Measure
Rules with Semantic Scores
Clustering
Centroid - Interesting Scores (I)

Figure 7. Transformer-Based Rule Processing

with a vocabulary size of about 30,000 tokens. Each transformer block has 768 hidden layers and 12 self-attention heads, and 110M parameters for training.

6.2. Interesting Centroids

Interesting cluster centroids are identified by applying the k-nearest neighbour (KNN) algorithm on the average word embedding. Table 12 and 13 illustrate the interesting rules derived using the healthcare BERT models. The interesting centroids from CovidBERT are found to represent the following concise demonstrations of the rules:

- Symptoms = {Fever}
- Treatment = {Medicine prescribed with home quarantine advice}
- Age = { 2, 3, 10}
- Category = {ILI}

The BioClinicalBERT represent the rules that have the following summary illustrations:

- Symptoms = {Fever}
- Treatment = {Medicine prescribed with home quarantine advice, Admit to another Hospital}
- Age = { 20, 35 }
- Category = {ILI, SARI }

By comparing the rules from both models, the BioClinicalBERT has detailed rules that have more interesting facts. Whereas CovidBERT indicates the basic and frequent level of attributes only.

6.3. Rules with Semantic Score

The cosine similarity index computes the degree of similarity between two vectors in an inner product space. It determines if two vectors point in the same general direction by computing the cosine of the angle between them. Referring to Table 14 and 15, the semantic rules represent the most interesting ones to the decision makers compared with the centroid method. The semantic rules illustrate the broad level of information like:

- Symptom = {Acute Febrile illness, Diabetic, Hypertension, Breathlessness, SARI

Table 12. Rules from cluster Centroid- CovidBERT

Cluster Centroid	Embedding Score	Rules
870	-1.62449046e-02	hasage 2 sufferfrom fever hascategory ILI treatmentprovided medicines prescribed with home quarantine advice
369	-1.61534358e-02	hasage 3 hascategory ILI treatmentprovided medicines prescribed with home quarantine advice'
620	-1.60850203e-02	hascategory ILI treatmentprovided referred to another hospital with call to emergency 108 or private ambulance livesin bangalore urban'
122	-1.61678973e-02	livesin bagalkot treatmentprovided medicines prescribed with home quarantine advice
1118	-1.64708573e-02	hasAge 10 sufferfrom fever hascategory ILItreatmentprovided medicines prescribed with home quarantine advice

Table 13. Rules from cluster Centroid - BioClinicalBERT

Cluster Centroid	Embedding Score	Rules
623	-6.75325135e-03	treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever livesin bangalore urban
873	-6.81287221e-03	hasage 20 hascategory ILI sufferfrom diabetic treatmentprovided medicines prescribed with home quarantine advice
123	-6.83343405e-03	livesin bagalkot sufferfrom SARI treatmentprovided medicines admittoanotherhospital
1119	-6.77445897e-03	sufferfrom fever treatmentprovided covidtest livesin chamarajanagar hascategory covidsuspect
372	-6.84563332e-03	hasage 35 sufferfrom hypertension treatmentprovided medicines prescribed with home quarantine advice

}

- Treatment = {Admitted to own hospital, Home quarantine advice, medicines prescription for the symptoms }
- Age = {55, 70, 43}
- Category = {ILI, SARI, COVID-19}

Table 14. Rules from CovidBERT with Semantic Scores

Cluster Centroids	Semantic Score	Rules
540630	0.73803685	livesin bagalkot sufferfrom fever⇒hasage 36 treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever hascategory ILI
76555	0.34837287	livesin bangalore urban treatmentprovided call to emergency 108 for covid 19 testing⇒livesin gadag hascategory ILI sufferfrom fever
385876	0.63081936	hasage 5 treatmentprovided medicines prescribed with home quarantine advice hascategory ILI ⇒sufferfrom afi hascategory ILI treatmentprovided medicines prescribed with home quarantine advice livesin bangalore urban
1119	0.86556687	treatmentprovided admitted to own hospital sufferfrom fever hascategory ILI⇒hascategory ILI livesin mysore treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever
694157	0.51231119	hasage 55 treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever⇒sufferfrom AFI hascategory ILI treatmentprovided medicines prescribed with home quarantine advice

Table 15. Rules from BioClinicalBERT with Semantic Scores

Cluster Centroids	Semantic Score	Rules
384851	0.93546483	livesin koppal sufferfrom diabetic hascategory ILI⇒livesin tumkur hascategory ILI treatmentprovided medicines prescribed with home quarantine advice
694510	0.98074915	livesin chitradurga treatmentprovided medicines prescribed with home quarantine advice hascategory ILI⇒hasage 55 treatmentprovided medicines prescribed with home quarantine advice sufferfrom breathlessness hascategory sari
76468	0.86965918	livesin yadgir sufferfrom diabetic ⇒treatmentprovided admittedtoownhospital sufferfrom fever livesin belgaum hascategory covid 19 suspect
540612	0.96043468	hasage 70 hascategory ILI treatmentprovided medicines prescribed with home quarantine advice⇒livesin koppal treatmentprovided medicines prescribed with home quarantine advice hascategory ILI sufferfrom fever
229887	0.91403103	hasage 43 treatmentprovided medicines prescribed with home quarantine advice hascategory SARI ⇒livesin bangalore urban sufferfrom breathlessness hascategory sari treatmentprovided admitted to own hospital

The summary of the semantic interesting rules covers the larger scope for the decision-makers. Also, by comparing the BERT models' rules, it's observed that BioClinicalBERT has semantic-rich information compared to the CovidBERT model. The similarity index is high with the rules about the SARI and breathlessness. Also, diabetes has high relevance with ILI patients.

6.4. Rules using Distance Measure

Introducing the distance-based measure for having interesting rules is widely seen in the literature. We have incorporated the distance-based measure for the semantic rules generated using the cosine similarity measure. The results tabulated in Table 16, 17 from CovidBERT and Table 18, 19 from BioClinicalBERT models represent the most interesting rules from the used COVID-19 corpora. The tables compare two rules from the input rule set and represent the distance between the two rules, highlighting the importance and relevance of the generated rules. The distance-based measure is applied to the five clusters generated by applying K-means clustering discussed in Section 6.2. These inferences also help to identify semantically similar rules for decision-makers. For instance, referring to Table 18 rule 1 illustrates the treatment provided relevance and rules 2 indicates the relevance between the patients.

7. Rule Evaluation

An effective rule evaluation framework is proposed to indicate the level of rule's interestingness. The framework has two dimensions: Statistical analysis and domain expert's evaluation.

7.1. Statistical Significance

In data mining, one well-studied strategy is inferring dependencies and interesting facts from data. Using Ontology to achieve the same goal will improve the semantic richness of inferred rules. This section describes the statistical method for evaluating

Table 16. Semantic Rules with Absolute Distance Measure using CovidBERT for Cluster 0,1 and 2

Rule1	Rule2	Semantic Score	cluster	Centroid	Absolute Dist.
hasage 5 treatmentprovided medicines prescribed with home quarantine advice hascategory ili.	livesin bangalore urban sufferfrom fever treatmentprovided admitted to own hospital.	0.5148	0	0.514836	3.61E-05
residesat belgaum sufferfrom sweating headache muscle aches loss of appetite dehydration general weakness sneezing and an itchy runny or blocked nose.	treatmentprovided call to emergency 108 for covid 19 testing hascategory covid 19 suspect sufferfrom fever livesin bangalore urban.	0.5148	0	0.514836	3.61E-05
hasage 20 sufferfrom fever treatmentprovided medicines prescribed with home quarantine advice.	livesin dharwad treatmentprovided medicines prescribed with home quarantine advice hascategory ili.	0.5148	0	0.514836	3.61E-05
livesin udupi treatmentprovided admitted to own hospital.	treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever hascategory ili livesin koppal.	0.5148	0	0.514836	3.61E-05
hascategory sari livesin mysore.	livesin bangalore urban sufferfrom fever treatmentprovided medicines prescribed with home quarantine advice hascategory ili.	0.5148	0	0.514836	3.61E-05
hasage 3 sufferfrom fever hascategory ili.	livesin raichur hascategory ili treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever.	0.7405	1	0.740541	4.14E-05
sufferfrom fever treatmentprovided medicines prescribed with home quarantine advice livesin udupi.	livesin chikkaballapura treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever hascategory ili.	0.7405	1	0.740541	4.14E-05
residesat chikmagalur sufferfrom sweating headache muscle aches loss of appetite dehydration general weakness.	treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever livesin bangalore urban.	0.7405	1	0.740541	4.14E-05
hasage 48 sufferfrom fever hascategory ili treatmentprovided medicines prescribed with home quarantine advice.	treatmentprovided admitted to own hospital sufferfrom fever hascategory sari livesin bangalore urban.	0.7405	1	0.740541	4.14E-05
hasage 2 treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever.	treatmentprovided admitted to own hospital sufferfrom fever livesin bangalore urban hascategory sari.	0.7405	1	0.740541	4.14E-05
treatmentprovided referred to another hospital with call to emergency 108 or private ambulance livesin belgaum.	hasage 54 treatmentprovided medicines prescribed with home quarantine advice hascategory ili.	0.6338	2	0.633753	4.72E-05
sufferfrom fever treatmentprovided call to emergency 108 for covid 19 testing.	hasage 33 treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever.	0.6338	2	0.633753	4.72E-05
livesin mysore treatmentprovided admitted to own hospital hascategory sari.	livesin dharwad sufferfrom fever hascategory ili treatmentprovided medicines prescribed with home quarantine advice.	0.6338	2	0.633753	4.72E-05
hasage 0 sufferfrom fever hascategory ili.	hasage 25 treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever hascategory ili.	0.6338	2	0.633753	4.72E-05
hasage 17 hascategory ili treatmentprovided medicines prescribed with home quarantine advice.	hascategory ili treatmentprovided medicines prescribed with home quarantine advice livesin bangalore urban.	0.6338	2	0.633753	4.72E-05

Table 17. Semantic Rules with Absolute Distance Measure using CovidBERT for Cluster 3 and 4

Rule1	Rule2	Semantic Score	cluster	Centroid	Absolute Dist.
hasage 58 hascategory ili treatmentprovided medicines prescribed with home quarantine advice.	hasage 7 hascategory ili treatmentprovided medicines prescribed with home quarantine advice.	0.8671	3	0.867076	2.43E-05
hasage 32 sufferfrom fever hascategory ili treatmentprovided medicines prescribed with home quarantine advice.	hascategory ili sufferfrom fever livesin tumkur treatmentprovided medicines prescribed with home quarantine advice.	0.8671	3	0.867076	2.43E-05
treatmentprovided medicines prescribed with home quarantine advice livesin ramanagara hascategory ili sufferfrom fever.	livesin bangalore rural sufferfrom influenza like illness.	0.8671	3	0.867076	2.43E-05
livesin udupi hascategory ili sufferfrom fever.	livesin shimoga treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever.	0.8671	3	0.867076	2.43E-05
treatmentprovided referred to another hospital with call to emergency 108 or private ambulance hascategory sari sufferfrom fever.	hasage 55 treatmentprovided medicines prescribed with home quarantine advice hascategory ili sufferfrom fever.	0.8671	3	0.867076	2.43E-05
livesin chikkaballapura treatmentprovided medicines prescribed with home quarantine advice.	hascategory ili livesin kalaburagi sufferfrom fever.	0.35	4	0.350039	3.88E-05
sufferfrom nothing treatmentprovided medicines prescribed with home quarantine advice.	treatmentprovided medicines prescribed with home quarantine advice livesin bangalore urban hascategory ili sufferfrom fever.	0.35	4	0.350039	3.88E-05
hasage 60 hascategory ili treatmentprovided medicines prescribed with home quarantine advice.	treatmentprovided call to emergency 108 for covid 19 testing livesin bangalore urban sufferfrom fever.	0.35	4	0.350039	3.88E-05
hasage 38 treatmentprovided medicines prescribed with home quarantine advice hascategory ili.	livesin kalaburagi hascategory ili sufferfrom fever.	0.35	4	0.350039	3.88E-05
residesat haveri sufferfrom sweating headache muscle aches loss of appetite dehydration general weakness.	reason 27 june trace history absent status c hasgender male.	0.35	4	0.350039	3.88E-05

Table 18. Semantic Rules with Absolute Distance Measure using BioClinicalBERT for Cluster 0,1 and 2

Rule1	Rule2	Semantic Score	cluster	Centroid	Absolute Dist.
hasage 4 hascategory ili treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever.	sufferfrom afi livesin bangalore urban treatmentprovided medicines prescribed with home quarantine advice hascategory ili.	0.9808	0	0.980824	2.42E-05
hasage 24 sufferfrom fever treatmentprovided medicines prescribed with home quarantine advice.	hasage 6 hascategory ili treatmentprovided medicines prescribed with home quarantine advice.	0.9808	0	0.980824	2.42E-05
hasage 21 hascategory ili sufferfrom fever treatmentprovided medicines prescribed with home quarantine advice.	hasage 3 hascategory ili treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever.	0.9808	0	0.980824	2.42E-05
livesin gadag treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever hascategory ili.	livesin koppal treatmentprovided medicines prescribed with home quarantine advice hascategory ili sufferfrom fever.	0.9808	0	0.980824	2.42E-05
hasage 34 treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever hascategory ili.	hasage 35 livesin bangalore urban treatmentprovided medicines prescribed with home quarantine advice hascategory ili.	0.9808	0	0.980824	2.42E-05
hascategory ili livesin kalaburagi sufferfrom fever.	reason 27 june trace history absent status c hasgender male.	0.871	1	0.870988	1.24E-05
sufferfrom for covid test treatmentprovided call to emergency 108 for covid 19 testing hascategory covid 19 suspect.	livesin gadag hascategory ili sufferfrom fever.	0.871	1	0.870988	1.24E-05
sufferfrom sweating headache muscle aches loss of appetite dehydration general weakness sneezing and an itchy runny or blocked nose prescribedwith fever drugs allergy drugs.	residesat koppal prescribedwith cough syrup.	0.871	1	0.870988	1.24E-05
residesat belgaum sufferfrom sweating headache muscle aches loss of appetite dehydration general weakness sneezing and an itchy runny or blocked nose.	hasage 26 sufferfrom fever treatmentprovided medicines prescribed with home quarantine advice.	0.871	1	0.870988	1.24E-05
residesat dharwad sufferfrom sweating headache muscle aches loss of appetite dehydration general weakness sneezing and an itchy runny or blocked nose.	hasage 33 treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever hascategory ili.	0.871	1	0.870988	1.24E-05
hascategory sari livesin mysore.	livesin haveri hascategory ili sufferfrom fever.	0.9365	2	0.936467	3.26E-05
hasage 39 hascategory ili treatmentprovided medicines prescribed with home quarantine advice.	treatmentprovided admitted to own hospital sufferfrom fever livesin mysore.	0.9365	2	0.936467	3.26E-05
hascategory sari livesin mysore.	hasage 30 sufferfrom fever hascategory ili treatmentprovided medicines prescribed with home quarantine advice.	0.9365	2	0.936467	3.26E-05
treatmentprovided referred to another hospital with call to emergency 108 or private ambulance hascategory ili.	treatmentprovided admitted to own hospital sufferfrom fever livesin bangalore urban hascategory sari.	0.9365	2	0.936467	3.26E-05
hascategory sari livesin bangalore urban.	hasage 31 hascategory ili treatmentprovided medicines prescribed with home quarantine advice.	0.9365	2	0.936467	3.26E-05

Table 19. Semantic Rules with Absolute Distance Measure using BioClinicalBERT for Cluster 3 and 4

Rule1	Rule2	Semantic Score	cluster	Centroid	Absolute Dist.
livesin mysore treatmentprovided admitted to own hospital.	hasage 25 sufferfrom fever treatmentprovided medicines prescribed with home quarantine advice.	0.9152	3	0.915191	9.43E-06
sufferfrom kodagu hascategory ili.	hasage 47 treatmentprovided medicines prescribed with home quarantine advice hascategory ili.	0.9152	3	0.915191	9.43E-06
livesin belgaum treatmentprovided call to emergency 108 for covid 19 testing.	hasage 42 treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever hascategory ili.	0.9152	3	0.915191	9.43E-06
hascategory ili treatmentprovided referred to another hospital with call to emergency 108 or private ambulance sufferfrom fever.	sufferfrom fever livesin shimoga hascategory ili.	0.9152	3	0.915191	9.43E-06
sufferfrom kodagu hascategory ili.	hasage 16 treatmentprovided medicines prescribed with home quarantine advice hascategory ili.	0.9152	3	0.915191	9.43E-06
hasage 5 hascategory ili treatmentprovided medicines prescribed with home quarantine advice.	hascategory ili livesin koppal treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever.	0.961	4	0.961009	9.46E-06
hasage 60 treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever.	sufferfrom afi treatmentprovided medicines prescribed with home quarantine advice livesin bangalore urban hascategory ili.	0.961	4	0.961009	9.46E-06
livesin bangalore urban treatmentprovided admitted to own hospital hascategory ili.	sufferfrom afi treatmentprovided medicines prescribed with home quarantine advice livesin bangalore urban hascategory ili.	0.961	4	0.961009	9.46E-06
sufferfrom fever livesin dakshina kannada hascategory ili.	livesin belgaum treatmentprovided medicines prescribed with home quarantine advice hascategory ili sufferfrom fever.	0.961	4	0.961009	9.46E-06
sufferfrom fever livesin dakshina kannada hascategory ili.	hascategory ili livesin udupi treatmentprovided medicines prescribed with home quarantine advice sufferfrom fever.	0.961	4	0.961009	9.46E-06

the inferred rule. To begin, we analyze the inferred rule using the Chi-square measure and decide on the defined hypothesis using the cross tab of inferred rules. Second, we consult domain experts about the top rules' relevance, significance, and use. The inferred rules are ranked based on the inputs of the domain expert to indicate their level of interest.

7.1.1. Chi-Square Measure

The inferred interesting rules are statistically significant according to the results tabulated in tables 20 and 21. We used Chi-Square (*chi*) to test the importance of the interesting rules. The p-value was found to be larger than 0.05 in all situations. The cross-tab representation refers to the antecedent and related consequent of a rule and its frequency of occurrence in both COVID-19 corpora. A lower p-value implies statistical significance for the null hypothesis. The p-value represents the likelihood of discovering these results if the null hypothesis is true. In this analysis, we take the statistically significant $p < 0.5$ into account.

Referring to Table 20, The chi-square statistic for the KATrace dataset is 38.6552. The p-value is .001219. The result is significant at $p < .05$.

Referring to Table 21, The chi-square statistic for the COKPME COVID-19 dataset is 27.7297. The p-value is .001058. The result is significant at $p < .05$.

Table 20. Chi square tabulation for KATrace sample data

	Covid-19	COVID-19 Susp.	MP-HQ	SARI	Cured
sCC	10 (8.53) [0.25]	12 (8.30) [1.65]	1 (5.57) [3.75]	1 (4.43) [2.66]	10 (7.16) [1.12]
hasDiagnosedFor	15 (12.29) [0.60]	17 (11.96) [2.12]	1 (8.03) [6.15]	1 (6.39) [4.55]	15 (10.32) [2.12]
ILI	20 (21.57) [0.11]	21 (21.00) [0.00]	18 (14.09) [1.08]	12 (11.22) [0.05]	15 (18.12) [0.54]
hasDiagnoisedFor	10 (10.79) [0.06]	8 (10.50) [0.59]	10 (7.05) [1.24]	10 (5.61) [3.44]	5 (9.06) [1.82]
Male, Female	20 (21.82) [0.15]	15 (21.24) [1.83]	19 (14.26) [1.58]	15 (11.35) [1.18]	18 (18.33) [0.01]

MP-HQ: Medicine Prescribed with Home Quarantine, ILI : Influenza-like Illness, sCC : sufferfromComorbidOthers, NCC: No Comorbid Conditions, COVID-19 Susp. : COVID-19 (Suspect).

Table 21. Chi square tabulation for COKPME sample data

	ILI, Breathlessness	Fever, Cold	Diabetic, Fever	OtherComorbidity
hasdiagnosedFor	24 (17.2) [7.32]	8 (9.91) [7.25]	12 (9.72) [4.90]	10 (8.60) [0.94]
medicineFor	22 (22.93) [0.09]	18 (17.44) [0.45]	8 (12.11) [0.08]	7 (9.24) [0.03]
SufferFrom	10 (6.74) [1.91]	8 (7.57) [0.07]	8 (2.87) [2.01]	6 (3.83) [1.24]
C0VID-19 Suspect	16 (6.60) [0.05]	18 (9.40) [8.80]	5 (7.08) [4.77]	3 (2.40) [0.03]

7.1.2. Domain Expert Evaluation

To have a potential method to use domain experts for rule evaluation, measuring the level of interestingness is proposed on a scale of one to five. Details are illustrated in Table 22. Two domain experts carried out the evaluation. The more important rules are thought to be more interesting rules. We ask the domain expert to rate a hypothesis based on its importance, relevance, and usefulness. The following Hypothesis cloud is considered:

(1) Rules that are with sufficient confidence than experts knowledge.
(2) Rules and experts' knowledge have the same phase or confidence.
(3) Rules with a lower level of confidence than expert's knowledge.

Considering these three hypotheses, rules are categorized and rated on a scale of one to five to identify the level of interestingness.

7.2. Implication and limitations of our Framework

The propagation of interestingness through ontology-based methods has been a new area of research in recent years. As in the healthcare domain, it mainly focuses on symptom analysis, diagnosis, and analyzing the consequences of the spread, symptom patterns, etc. As a result, it is critical to find interesting facts that could help the decision-makers. The interesting patterns lie as implicit facts in the data. With domain ontology, it can be uncovered with semantic knowledge. However, this is only the first step toward detecting interestingness in data using domain ontology, and there is always an opportunity for improvement. Our Semantic Interesting Framework (SIF) has some shortcomings that will be addressed in future work. Some of the limitations

Table 22. Evaluation Scale Range

Range	Illustration	Remarks
1	Irrelevant	Rules do not make sense to the defined hypothesis.
2	Low-level	Rule has a low level of significance to the hypothesis
3	Mid-level	Rule has mid-level of significance to the hypothesis
4	High-level	Rule has a high level of significance to the hypothesis
5	Very High-level	Rule has a very high level of significance to the hypothesis

are as follows.

- The proposed framework uses the pre-trained BERT model for rule semantic scores. However, efforts can be made to use the distance-based similarity measure.
- Our proposed framework is slightly biased towards the user interest, such as using the constraint file defined by the user keeping ontology as a reference.

8. Conclusion

This study makes a significant contribution by mining interesting patterns from RDF data using instance and schema-level information. Most current research focuses on mining instance-level data to uncover interesting relationship rules. To solve this issue, we present a SIF, a unique methodology that uses knowledge encoded at the schema level through an ontology's relationship and *ConstApriori* for instance level rules. The semantically-enriched rules are generated using the schema level relations like *rdf:type and rdfs:subClassOf.* The rules with their semantic scores are generated using the transformer-based method.

Further, COVID BERT and clinical BERT, and Bio Bert Models are used to find the most interesting rules using the equivalence of variance method. It showed that the variance is equally distributed among all the interesting clusters. Finally, as an evaluation criterion, we use the Chi-square test for significance and domain expert evaluation mechanism to assess the generated rules. It found that the semantic rules from SIF are significant at a p-value of 0.95. Interestingly, the transformer-model identified and domain experts ranked rules have a high level of correlation.

References

Afolabi, Ibukun, Olaperi Sowunmi, and Olawande Daramola. 2017. "Semantic association rule mining in text using domain ontology." *International Journal of Metadata, Semantics and Ontologies* 12 (1): 28–34.

Agrawal, Rakesh, Ramakrishnan Srikant, et al. 1994. "Fast algorithms for mining association rules." In *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215, 487–499. Citeseer.

AL-Zawaidah, Farah Hanna, Yosef Hasan Jbara, and AL Marwan. 2011. "An improved algorithm for mining association rules in large databases." *World of Computer science and information technology journal* 1 (7): 311–316.

Alzubi, Jafar A, Rachna Jain, Anubhav Singh, Pritee Parwekar, and Meenu Gupta. 2021. "COBERT: COVID-19 question answering system using BERT." *Arabian journal for science and engineering* 1–11.

Apostolopoulos, Ioannis D, and Tzani A Mpesiana. 2020. "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks." *Physical and engineering sciences in medicine* 43 (2): 635–640.

Arora, Parul, Himanshu Kumar, and Bijaya Ketan Panigrahi. 2020. "Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India." *Chaos, Solitons & Fractals* 139: 110017.

Badenes-Olmedo, Carlos, David Chaves-Fraga, MarÍa Poveda-VillalÓn, Ana Iglesias-Molina, Pablo Calleja, Socorro Bernardos, Patricia MartÍn-Chozas, et al. 2020. "Drugs4Covid: Drug-driven Knowledge Exploitation based on Scientific Publications." *arXiv preprint arXiv:2012.01953* .

Barati, Molood, Quan Bai, and Qing Liu. 2016. "SWARM: an approach for mining semantic association rules from semantic web data." In *Pacific rim international conference on artificial intelligence*, 30–43. Springer.

Bellandi, Andrea, Barbara Furletti, Valerio Grossi, and Andrea Romei. 2007. "Ontology-driven association rule extraction: A case study." *Contexts and Ontologies Representation and Reasoning* 10.

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The semantic web." *Scientific american* 284 (5): 34–43.

Biradar, Shankar, Sunil Saumya, and Arun Chauhan. 2022. "Combating the infodemic: COVID-19 induced fake news recognition in social media networks." *Complex & Intelligent Systems* 1–13.

Bizer, Christian, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. "Dbpedia-a crystallization point for the web of data." *Journal of web semantics* 7 (3): 154–165.

Bringmann, Björn, Siegfried Nijssen, and Albrecht Zimmermann. 2011. "Pattern-based classification: A unifying perspective." *arXiv preprint arXiv:1111.6191* .

C, Abhilash, and Kavi Mahesh. 2021. "Graph Analytics Applied to COVID19 Karnataka State Dataset." In *2021 The 4th International Conference on Information Science and Systems*, New York, NY, USA, 74–80. Association for Computing Machinery. https://doi.org/10.1145/3459955.3460603.

Choi, Eunsol, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. "QuAC: Question answering in context." *arXiv preprint arXiv:1808.07036* .

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

dos Santos, Fabiano Fernandes, Marcos Aurelio Domingues, Camila Vaccari Sundermann, Veronica Oliveira de Carvalho, Maria Fernanda Moura, and Solange Oliveira Rezende. 2018. "Latent association rule cluster based model to extract topics for classification and recommendation applications." *Expert Systems with Applications* 112: 34–60.

Dutta, Biswanath, and Michael DeBellis. 2020. "CODO: an ontology for collection and analysis of COVID-19 data." *arXiv preprint arXiv:2009.01210* .

Garbe, Wolf. 2012. "SymSpell." 6. hhttps://github.com/wolfgarbe/SymSpell.

Garijo, Daniel. 2017. "WIDOCO: a wizard for documenting ontologies." In *International Semantic Web Conference*, 94–102. Springer.

Geng, Liqiang, and Howard J Hamilton. 2006. "Interestingness measures for data mining: A survey." *ACM Computing Surveys (CSUR)* 38 (3): 9–es.

Guo, Xiao, Hengameh Mirzaalian, Ekraam Sabir, Ayush Jaiswal, and Wael Abd-Almageed. 2020. "Cord19sts: Covid-19 semantic textual similarity dataset." *arXiv preprint arXiv:2007.02461* .

He, Yongqun, Hong Yu, Edison Ong, Yang Wang, Yingtong Liu, Anthony Huffman, Hsin-hui Huang, et al. 2020. "CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis." *Scientific data* 7 (1): 1–5.

Manda, Prashanti, Fiona McCarthy, and Susan M Bridges. 2013. "Interestingness measures and strategies for mining multi-ontology multi-level association rules from gene ontology annotations for the discovery of new GO relationships." *Journal of biomedical informatics* 46 (5): 849–856.

Mangla, Monika, and Rakhi Akhare. 2015. "Association rules filtration using dynamic methods." *International Research Journal of Engineering and Technology* 2 (3): 1103–1106.

Marinica, Claudia, and Fabrice Guillet. 2010. "Knowledge-based interactive postmining of association rules using ontologies." *IEEE Transactions on knowledge and data engineering* 22 (6): 784–797.

Moreno, María N, Saddys Segrera, and Vivian F López. 2005. "Association Rules: Problems, solutions and new applications." *Actas del III Taller Nacional de Minería de Datos y Aprendizaje, Tamida* 317–323.

Musen, Mark A. 2015. “The protégé project: a look back and a look forward.” *AI Matters* 1 (4): 4–12. https://doi.org/10.1145/2757001.2757003.

Ozturk, Tulin, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, and U Rajendra Acharya. 2020. “Automated detection of COVID-19 cases using deep neural networks with X-ray images.” *Computers in biology and medicine* 121: 103792.

Qin, Lei, Qiang Sun, Yidan Wang, Ke-Fei Wu, Mingchih Chen, Ben-Chang Shia, and Szu-Yuan Wu. 2020. “Prediction of number of cases of 2019 novel coronavirus (COVID-19) using social media search index.” *International journal of environmental research and public health* 17 (7): 2365.

Schuster, Mike, and Kaisuke Nakajima. 2012. “Japanese and korean voice search.” In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5149–5152. IEEE.

Shan, Guohou, Lina Zhou, and Dongsong Zhang. 2021. “From conflicts and confusion to doubts: Examining review inconsistency for fake review detection.” *Decision Support Systems* 144: 113513.

Shawe-Taylor, John, Nello Cristianini, et al. 2004. *Kernel methods for pattern analysis.* Cambridge university press.

Shen, Iris, Le Zhang, Jianxun Lian, Chieh-Han Wu, Miguel Gonzalez Fierro, Andreas Argyriou, and Tao Wu. 2020. “In search for a cure: recommendation with knowledge graph on CORD-19.” In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3519–3520.

Srikant, Ramakrishnan, and Rakesh Agrawal. 1995. “Mining generalized association rules.” .

Suchanek, Fabian M, Gjergji Kasneci, and Gerhard Weikum. 2007. “Yago: a core of semantic knowledge.” In *Proceedings of the 16th international conference on World Wide Web*, 697–706.

Tandan, Meera, Yogesh Acharya, Suresh Pokharel, and Mohan Timilsina. 2021. “Discovering symptom patterns of COVID-19 patients using association rule mining.” *Computers in biology and medicine* 131: 104249.

Tomar, Anuradha, and Neeraj Gupta. 2020. “Prediction for the spread of COVID-19 in India and effectiveness of preventive measures.” *Science of The Total Environment* 728: 138762.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention is all you need.” *Advances in neural information processing systems* 30.

Wikipedia contributors. 2021. “FAIR data — Wikipedia, The Free Encyclopedia.” [Online; accessed 24-August-2021], https://en.wikipedia.org/w/index.php?title=FAIR$_dataoldid = $ 1038845392.