https://doi.org/10.48048/tis.2021.47

Ontology Based Text Classifier for Information Extraction from Coronavirus Literature

M Sivakami[*] and M Thangaraj

Department of Computer Science School of Information Technology, Madurai Kamaraj University, Madurai, Tamilnadu, India

([*]Corresponding author's e-mail: sivakamimk@gmail.com)

Received: 20 October 2020, Revised: 7 May 2021, Accepted: 20 May 2021

Abstract

The world is fighting an unprecedented coronavirus pandemic, and no country was prepared for it. Understanding the nature of this disease, when there is no available cure, is vital to encourage accurate clinical diagnosis and drug discovery prospects. When the amount of literature available is vast, it is important to represent the disease domain as completely as possible. The system should capture the morphology, semantics, syntax, and pragmatics of the given literature, in order to extract useful information. Also, building a classifier for a particular domain suffers from a zero frequency issue. To solve this effectively, latent topics are extracted and semantically represented in ontology to build a text classifier for coronavirus literature. The classifier is equipped with 2 components- 'ontology' and 'machine learning data model'. Ontology helps to model the morphology and the semantic and pragmatic aspects of the text data through Latent Drichlet Allocation (LDA). It also preserves the contextual information in the document space, providing holistic feature representation facilities. To solve zero frequency and to extract actionable insights, a machine learning algorithm, Multi class Support Vector Machine (M-SVM), is incorporated with the ontology. It encodes features and achieves a classifier with highly discriminated classes. Further, to preserve contextual information space, and to enable data model formulation, the ontology is generated as a knowledge graph with their respective predefined classes. The resulting dataset can be used for clinical diagnosis and further research on the disease. Experimental results have shown that the proposed classifier outperforms the existing systems, with better domain representation.

Keywords: COVID-19, Multi class classifier, Text classification, Ontology, Knowledge graphs

Introduction

The world is coping with a coronavirus pandemic disease that started in the year 2019, known as COVID-19. The situation is more unfortunate, since there is no cure for this deadly novel virus. In a situation like this, whatever knowledge that is available about the disease is of great value. Though the disease started spreading in late 2019, it had already been studied for more than a few decades, and numerous research articles had been produced in this regard. These articles not only carried basic information related to the novel virus, but also specified constructive research findings in the most elaborate manner. It will be highly useful if actionable information could be retrieved from such literature. It helps to understand the severity of the illness in patients testing positive for the virus. Also, informative guesses can be made in cases where immediate testing is not possible, as cases are increasing every hour. Additionally, the drug discovery aspects could also be probed using the knowledge about the disease.

One problem with the available literature is that it is very vast and, in order to achieve a good information extraction data model, the disease domain has to be represented in a comprehensive manner. Therefore, such a system should be able to extract the morphology about the novel virus and its various semantic and pragmatic foundations, along with the syntactic nature of the text content. Another problem with domain based classifiers is zero frequency. This is where, a particular class representing an element of the domain may not have sufficient features. This is alleviated efficiently using a topic modeling algorithm, where latent topics are extracted and built as ontology.

Considering the benefits of having an ontological knowledge base, from its semantic representation of concepts, to identifying relationships between similar concepts, the application of ontology as a knowledge base for classifiers is under-researched. This is mainly due to a lack of domain knowledge to populate the ontology. In one of the existing works related to CODO ontology [26], only the statistical characteristics of COVID-19, data based on patients' geographical location, was studied. The data model, however, was built from the already annotated knowledge in the websites. The work was a mere representation of statistical information viable to be queried. The applicability of the model for extracting new knowledge was not discussed, which is the actual need of the hour, when there is no cure for the disease available. The semantic nature of concepts about the disease was also not considered, which may have led to crucial information loss, as two or more entities may seem similar, but refer to different classes. Further, the work did not discuss the applicability of the model for solving document classification problems. The aim of the study was only to build a vocabulary database for future projects. Hence, this work focuses on building a 2 tier classifier.

To avoid the zero frequency and information extraction issues, one handles the data representation using 'ontology', and the other is used for extracting actionable information from the data using 'machine learning data model'. Ontology helps to characterize the information contained in the domain without losing the structure of the content, its contextual subtleties, or the semantic and syntactic qualities of the text content of the literature. When using ontologies to represent the domain, the feature space is expanded to accommodate all legitimate information about the domain. To preserve the contextual information space, and to enable automated information extraction through machine learning, the ontology is generated as a knowledge graph with their respective predefined classes. This results in the dimensionality issue, leading to data overfitting.

Overfitting can again be solved using the machine learning classifier, which automates feature encoding and feature selection, irrespective of the size of feature space. In the proposed work, an M-SVM is used to automate information extraction from the coronavirus disease literature. M-SVM produces an effective hyperplane in separating multiple classes, which is a basic requirement for such classifiers. The experiments were carried out using COVID-19 research articles generated in the last 20 years. The literature was analyzed for extracting useful information, which was categorized accordingly. The resulting dataset can be used for clinical diagnosis and further research on the disease. Experimental results have shown that the proposed classifier outperforms the existing systems, and also better represents the COVID-19 domain.

The remainder of the paper is organized as follows: Section 2 delivers the various directions of research in this area. The proposed architecture, with elaborate discussion of the working of the model, is discussed in section 3. The system is compared, with benchmark classifiers to reason out the performance of the system, in section 4. Section 5 discusses the major research findings, and section 6 presents the conclusion and future enhancement.

Literature review

The basic idea behind the formulation of our work is information extraction from documents. Extracting useful information from research literature will be highly useful for practitioners of a domain in identifying functionality improvements. The literature produced in the field of medicine offers one such significant opportunity. However, due to the vastness of the collection, and the highly complex nature of the literature, the system calls for sophisticated data models. Clinical text mining applies text classification approaches to such documents to extract patient health status [1].

To improve classification accuracy, the feature extraction has to be done effectively. Abdollahi *et al*. [1] employed 2 step feature engineering for identifying generic and disease specific features from an i2b2 document dataset and built an ontology. The large feature space of the ontology arising out of the process was searched using the Particle Swarm Optimization algorithm to identify minimal feature subsets. The model, however, suffered from delay, due to embedding features as particles, and was also not able to scale to new data about Coronary Artery Disease. Gayathri and Kannan [8], similar to clinical text mining, obtained biomedical text mining extracts for deeper insights for the prevention and cure of various diseases. The domain ontology and semantic document descriptions were used to build a document classifier for classifying Ayruvedic documents. The semantic matching techniques and the K-Nearest Neighbor algorithm retrieved semantically richer results. The size of data used by the model was less, and also suffered from a data imbalance problem. Synonymy, Polysemy, and Multi-word concepts are the curse of medical literature [14]. Exact information can only be extracted using semantic matching between domain ontologies and concept maps from documents. The maps help to relate two or more

concepts and to assign weighting for similar concepts using a similarity matrix. The incorporation of semantic knowledge into a text classification framework enriched document classification for the medical domain, but thematic maps were required to be constructed every time.

Medical ontologies capture relationships between concepts from MeSH and MEDLINE ontology, which contains data about medical subject headings [4]. This work deals with document classification using MeSH ontology by extending existing document representation. The ontology helps to disambiguate medical texts by annotating conceptual references in literature. The MEDLINE and MeSH ontologies are compared for semantic similarity using edge count measure of nodes in both ontologies and aggregate those classes that contain similar edge counts. This method might sometimes lead to polysemy, as concepts could overlap in a literature collection. Another kind of ontology related to medicine domain studies about the statistical characteristics of COVID-19 data is based on patients' geographical location. The ontology was built as a data model to represent disease knowledge about COVID cases, such as statistics about the number of active, recovered, deceased, and migrated cases using the patients' geo location. The data was further extended with patient data regarding nationality, symptom, suspected disease level, facility of treatment available, relationship between patients, reason for transmission, test result characteristics, etc. [26].

Another approach deals with using thematic graphs from documents to classify them according to contexts mentioned in the domain ontology [3]. It assigns topics dynamically, eliminating the need for model retraining. It uses Wikipedia data as ontology to classify news categories. Using Wikipedia for news context representation will render the system imprecise, and also imbalanced in covering certain domains. A similar document classification problem is dealt with using the concept of term weighting for categorizing documents [7]. The similarity between documents and ontology is calculated using WordNet and ranked according to their similarity scores. The experiments were conducted using web documents in pre-defined categories. The model failed to analyze the word sense of contents while conducting classification.

Text classification finds an interesting application in the field of knowledge management [6]. The crucial part of knowledge management deals with filtering useful information from non-useful information. The process is semantically carried out using ontologies which eliminate synonymy problem of documents using conceptual representations. The similarity mapping engine represents the meaning of sentences with the help of syntactic elements, lexical intonations, and contextual meaning obtained through ontology. However, the model is tested only with small-sized contextual data, and its applicability for large data size is still unproven.

A software requirements ontology was built for tagging Non-Functional Requirements using gold standard corpus [12]. The ontology was further extended using a support vector machine classifier that classified SRS document collection into classes defined in the ontology. The classification happened at the sentence level. The misclassification rate was still higher from the results, which could be further improved using syntactic and semantic features combined. A similar problem that classified security requirements from other requirements in SRS was probed [11]. The existing domain dependent classifiers were not able to be applied to the security requirement classification of other domains. Therefore, an ontology based linguistic model was trained on a domain-independent security requirement classifier using machine learning classifiers. The linguistic rules were manually constructed to extract generic concepts related to security requirements, which consumed much time and necessitated the use of more sophisticated NLP techniques. Multi-classification of functional requirements in the accounting domain was proposed in this work [13]. Here, a bag-of-words was modeled into ontology to gain word order, as well as conceptual representations. The model checked for various functional requirements, but in a very limited dataset, which led to biased classification, and the model could not be generalized for other domains, requiring complete retraining.

Ontology based LDA model was implemented for extracting information related to transportation from social media networks [2]. The combination of word embedding and lexicon based techniques improved accuracy of the model. The transportation features were obtained with semantic knowledge of the ontology, and LDA assigned various topics related to transportation. Polarity was assigned using machine learning and deep learning models. The performance was tested using Root Mean Square Error and Mean Average Error metrics. However, the model did not remove irrelevant words before sentiment tagging; this resulted in imprecise classification. The extraction of transport related terms was required in the pre-processing stage.

For better decision making support during rare events like earthquakes, floods, etc., Ontology Based Statistical Relational Learning and Machine Learning are used [5]. However, predicting rare events suffers from relative (class-imbalance) and absolute (small sample size) rareness. The complexity

increases when explainability is added as another dimension. Instead of a blackbox classifier, a transparent classifier can be built using domain ontologies, ML classifiers, and Fuzzy probabilistic logic. Hawalah [9] built an Arabic Ontology based topic classification model for Arabic text classification. The model depended on simple TF-IDF and Cosine similarity algorithms for classification. The inclusion of semantic features with semantic clustering mechanism were proven to improve accuracy. Nevertheless, the size of ontology taken for the study was small, and only single level classes were used. Kim and Rhee [10] undertook the labeling of the most influential topics in articles published in scientific literature in the datamining field. The social network analysis techniques, along with domain ontology, were used to assign topics to various articles. This avoided the ambiguous interpretation of topics in topic modeling but failed to capture the relationships between topics. The proposed work is the application of our previous work, published in Thangaraj and Sivakami [15]. The OntoClassifier model is applied for information extraction from the coronavirus literature collection.

Some of the problems identified from the domain are:

The concepts related to the nature of disease, medications, and virus-types are spread all over the literature collection.

Labeling unstructured medical concepts is tedious, given the complex nature of short expressions.

The presence of non-domain related words after the pre-processing stage leads to biased classification results.

The non-availability of domain information affects semantic feature extraction in document classification problems.

Conventional clinical document classification problems suffer from zero frequency.

Proposed methodology

The proposed system is named the COVID-OntoClassifier (COVOC). It is a 2 step classifier, built for extracting useful information from the coronavirus literature. The aim is to extract data related to nature of the disease and its medicine prescriptions, area of infection, viral nature, genetic markup, etc. The topics are defined automatically, and corresponding topic related terms are extracted using similarity measures. These are the inherent concepts present in the literature that are capable of providing insights about the disease. The coronavirus research literature is obtained from the Coronavirus Open Research Dataset (CORD). The data is preprocessed, and various concept topics are identified in the topic modeling phase. The concepts are represented as ontology, along with their topics as classes. The ontology is embedded as a knowledge graph, with all its conceptual terms and target classes. This data is fed into the machine learning data model to effectively classify topic specific concepts and to extend the knowledge representation. The detailed process flow for the proposed classifier is given in **Figure 1**.

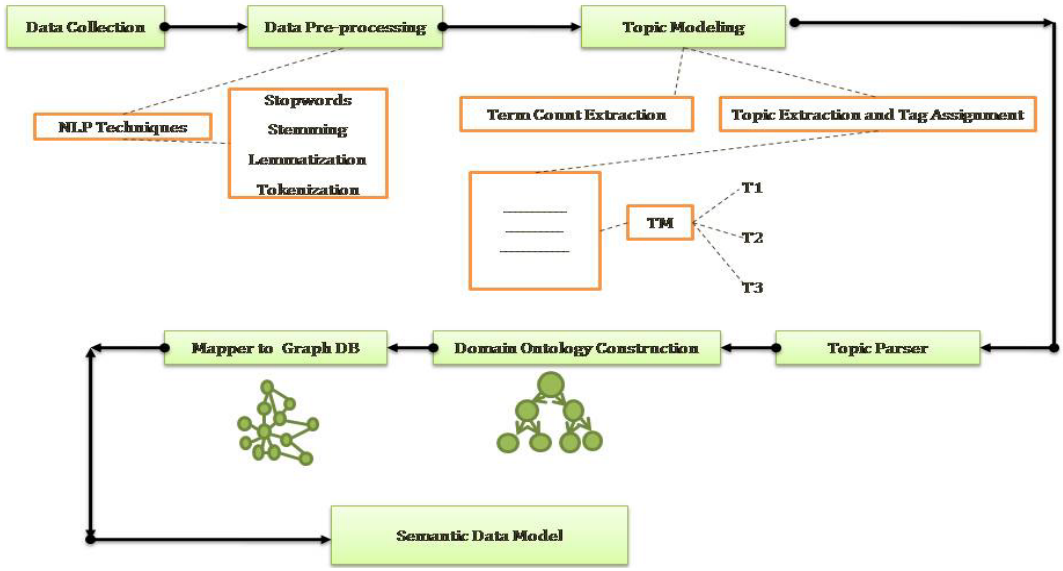

Figure 1 COVOC framework.

Data collection

The experiments were carried out using coronavirus research articles generated in the last 20 years. The literature was analyzed for extracting useful information and then categorized accordingly. The data was obtained from the open source resource called the COVID-19 Open Research Dataset (CORD-19). The dataset is a collection of research articles on viruses related to the coronavirus family. The data was extracted according to the requirements of our work and parsed into 8 separate columns, 'Name of the Source', 'title', 'DOI', 'Abstract', 'Date of Publication', 'Author Name', 'Journal', and 'pdfs as .json', amounting to a total of 1,12,708 records. From the initial Exploratory Data Analysis given in **Figure 2**, the distribution of words in the abstract was studied, and it was found that the average length of all abstracts was 200 words, with minor peak values on both sides of the mean.

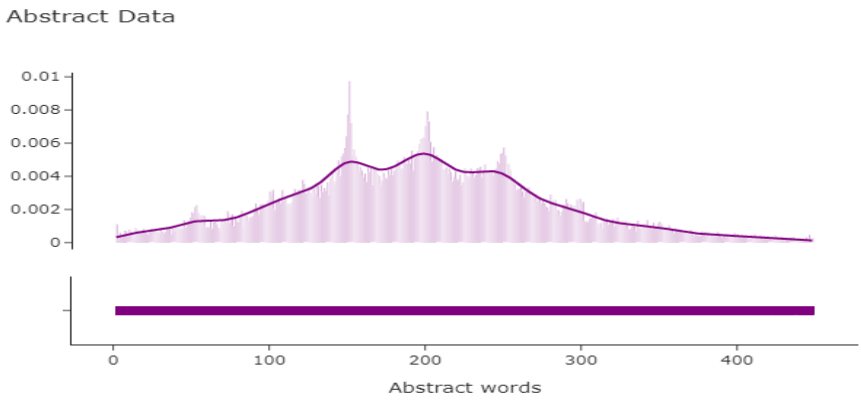

Figure 2 Distribution of words in extracted 'abstracts'.

Data preprocessing

The data was cleaned to remove noisy and unwanted data from the analysis. This phase was customized according to the disease dataset. Medical texts require exclusive techniques at each stage, due to the unique and complex nature of data. This helps to avoid classification errors and biased results. Punctuation and special characters were removed in the first stage, as they did not possess any semantic value. Numbers were not removed, since many of the medicine and disease names come with numeric labels. Removing stopwords decreased data dimensionality. General case conversion, stemming, and

lemmatization are carried out. Besides these, words unrelated to the medical domain were removed after comparing them with MeSH and UMLS Metathesaurus. This helped to obtain a rich vocabulary set for the coronavirus disease domain.

Topic modeling

Topic modeling is the process of identifying latent topics from data [16]. LDA is one of the topic modeling algorithms widely used in data mining. It is mostly used in scenarios where topics are not known prior. Drichlet is the distribution of words in topics and topics in documents. The word corpus, obtained from the previous step after pre-processing, is used for probabilistic modeling using LDA. We chose LDA for our problem, due to its ability to deal with long documents and the large vocabulary size of text data [17]. The existing LDA was slightly modified to combine 'relation', also known as 'properties', extraction, to better suit the needs of medical literature. Before proceeding with topic extraction, the 'hypernyms', or top level class words, were extracted from the word corpus using the Metathesaurus and indexed separately; this was to form hierarchical topics.

***Improved*-LDA topic extraction**

In order to develop the latent topics, documents were converted into 'n' words. The words were tagged with their respective Parts of Speech (PoS) tags to extract the entities that reflected 'relationship' types. The PoS tags used for our work were 'Coordinating Conjunction', 'Foreign Word', 'List Item Marker', 'Nouns & its family', 'Verbs & its family', 'Numeral/Cardinal', and 'Preposition'. The 'n' words were grouped according to their PoS nature, either as triplets, bi-grams, or unigrams.

Step 1: Poisson distribution of sets of 'n' words

Step 2: The topics are denoted by 'Θ' and Drichlet distribution selects the distribution per document and in the entire document corpus.

Step 3:
(1) Choose a topic related to coronavirus disease, say, $To^{(i)} \sim Multinomial^{(\Theta)}$.
(2) Choose a word from word set 'n', denoted as $Ws^{(i)}$ *from* $P(Ws^{(i)}|To^{(i)},\beta)$.

Marginal document distribution as obtained from the above:

$$P(d) = \int\left(\prod_{i=1}^{n}\sum_{To^i} P(Ws^i \mid To^i,\beta)P(To^i\,|\Theta)\right)P(\Theta|\alpha)d\Theta. \quad (1)$$

Drichlet distribution was obtained using the 'α' parameter from '$P(\Theta|\alpha)d\Theta$'. The 'β' parameterized number of word sets under each topic, as given in '$P(Ws^i \mid To^i,\beta)$'. With the help of 'α', the prior distribution of topics could be obtained before the words belonging to a topic were seen in each document. The same prior knowledge was obtained for distribution of word sets through hyperparameter 'β'. The probability distribution of topics over the entire document corpus could be obtained using the formula:

$$P(D) = \prod_{i=1}^{M} P(d). \quad (2)$$

The number of classes and other statistics related to topics obtained through our *Improved*- LDA is given in **Table 1**. 'Properties' and 'Children' were topic word sets obtained from the corpus and designated in semantic terms to facilitate further processing.

Table 1 Topic statistics obtained using LDA.

Classes	**4,243**
Individuals	1,382
Properties	111
Maximum depth	43
Maximum number of children	961
Average number of children	23
Classes with a single child	240
Classes with more than 25 children	89

Though the topic modeling step itself classified the concepts related to the disease data, the huge feature space formed became an issue. Hence, this stage could be considered as an intermediate classifier, which is to be further filtered to obtain more specific features. These features should also preserve the semantic variations in the text to aid in unbiased classification. This is a major issue in medical information extraction, as wrong diagnosis could cost lives. Further, it should also facilitate the seamless accommodation of new data in the near future. Also, storage issues, such as RAM insufficiency, could arise when the topic data is stored as bags-of-words or 'thematic maps'. To solve these, the topics obtained are parsed into an ontology to save access time and run time memory consumption.

Domain ontology

Constructing ontology through this method has proven to be more efficient than extracting text triples, which brings noisy data when data is entirely unstructured. The Topic Parser function populated the ontology with topic classes and their respective word sets. In the first level, the parent classes were formed from the hypernym classes obtained from the topic modeling phase, along with their respective word distributions. In the second level, all the children nodes and their respective sub nodes were populated. Once the topics were embedded in the ontology, the coronavirus domain ontology could be obtained with well-connected information about the nature of the disease. The other implicit relationships that exist in the nodes could be obtained using reasoners in ontology. The nodes with their properties became instances for further processing and also reduced the feature space and data dimensionality. This also provided for a wholesome knowledge representation to facilitate any further research in this area. To eliminate the need for laboriously searching the ontology every time, the COVOC ontology was obtained in RDF format and converted into a Graph Database. After embedding the ontology as a knowledge graph, a machine learning data model was introduced. This model learned the associations among the nodes through training and became a semantic data model for all further research related to the coronavirus domain, such as correlation extraction, classification, named entity extraction, document classification, etc. The class distribution of information related to COVID is presented in **Figure 3**.

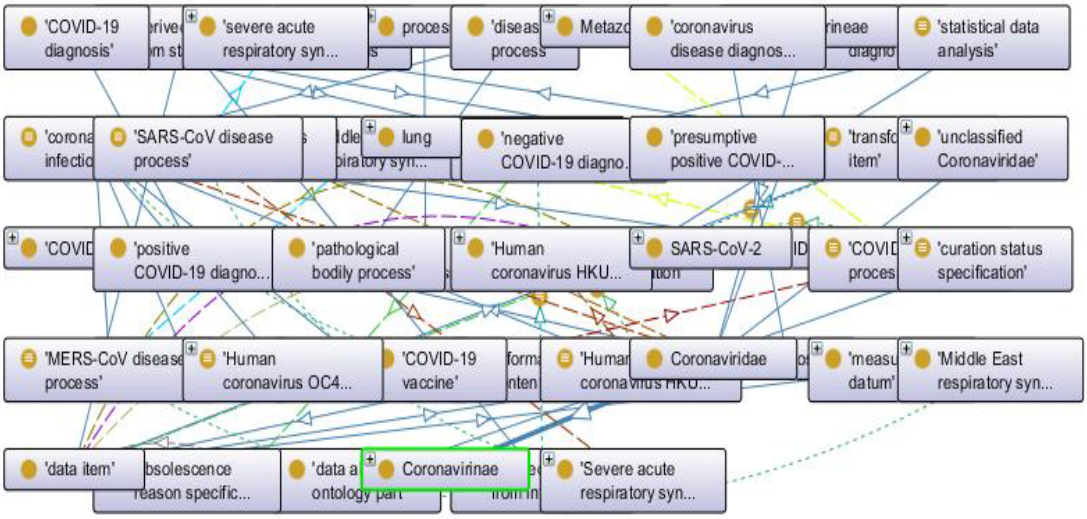

Figure 3 Class distribution of COVID-19 dataset.

Semantic data model

The data model was built to preserve the rich feature space and avoid lengthy tree traversals in a connected data like ours. For achieving this, the M-SVM algorithm was used, as it has minimum bias and sharper hyperplane separation between classes [18]. It also generalizes effortlessly along with the curse of dimensionality. In our work, the training dataset was given as Td $= \{x_{ijk\ldots} + x_i + x_j + x_k \ldots . x_{lmonp,} Cl_1, Cl_2, Cl_3 \ldots . . Cl_n\}$; *i to p* represented features given as input, and *Cl* represented class labels. In this M-SVM, the input variables were vectorized to high dimensional feature space to build an optimal separation hyperplane. Those vectors occupying the boundary of the

hyperplane were used for classifying unseen instances. The distance between the separating hyperplane and the closest data point belonging to each class is given in Eq. (3):

$$\text{sgn}\left(\sum_{i=1}^{n} y_i \beta_i . K\left(x_i, x_j\right) + b\right). \quad (3)$$

In this equation, x denotes the support vectors and coefficient values, and bias parameters are obtained using the Lagrange dual equation, as in Eq. (4):

$$\text{Max}\left(\sum_{i=1}^{n} \beta_i - 1/2 \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_i \beta_j . y_i y_j . K\left(x_i, x_j\right)\right). \quad (4)$$

where it is to be noted that,

$$\sum_{i=1}^{n} \beta_i y_i = 0. \quad (5)$$

Polynomial kernel function is used to encode dataset into hyperplanes by converting them to linear kernels. When the parameter $'\beta_i'$ varies between o and regularization factor 'C', the $'x_i'$ becomes a support vector. The regularization factor 'C' measures the swapping between margin and error, thereby minimizing the model's complexity. The overfitting of the model can be easily identified with higher values of 'C', as the M-SVM penalizes points that are non-separable. Also, smaller 'C' values denote a heavily underfitted model, which is also undesirable. The parameter settings play a crucial role in the overall results of the model; hence, it is imperative to have concrete settings for vital parameters. The most vital parameters are 'C', depended on Lagrangian multipliers, kernel function, tolerance margin, etc. The model parameters can be tuned using cross validation technique, which helps to select best values for each parameter in each run. When using SVM for multiclass, it suffers from rejection area dilemma in one vs rest classification. In one-vs -rest model, the classifier is trained for each class separately, treating examples that belong to a class as positives, and everything else as negatives. When there exists only one decision function, some of the examples during each class training might fall continuously in the rejection area. To avoid such a deadlock, the distance between rejected examples and decision function is calculated using the formula:

$$F_i(E^*) = \frac{|D_i(E^*)|}{\omega_i}. \quad (6)$$

ω_i denotes the normal support vector for the i[th] best hyperplane, and the best hyperplane can be constructed for decision function D, given as:

$$DF_i(E^*) = \sum_{z=1}^{n_{Fv}^i} \varphi_{iF} \acute{y}_{iF} K(E_{iF}. E^*) + b^i. \quad (7)$$

n_{Fv}^i signifies the number of support vectors gained, φ_{iF} is the Lagrange multiplier, Fv is the optimal hyper plane derived, and $K(E_{iF}. E^*)$ denotes the kernel function. An example should be classified into the class with longer distance, to avoid rejection dilemma.

The resulting COVOC classifier had a multitude of features, amounting to more than 10,000 COVID related features. A sample of the dataset obtained using the algorithm is given in **Table 2**. This dataset can be used to analyze correlation between various attributes to extract the most commonly used drugs, classification of crucial information, similarity between virus families, relation between human anatomy and virus RNA, infection levels, etc.

Table 2 Sample of significant concept classification from coronavirus literature.

COV	Disease	Infection	Antimalarial	Virus
covs	schistosomiasis	virus-associated	amodiaquine	pneumoviridae
coronaviruses	endocarditis	Superinfection	artemisinin	cpiv
sars-covs	food-and	Modestly	artesunate	pneumovirus
sl-cov	reemergence	Reflecting	antiparasitic	ross
merbecovirus	leptospirosis	post-viral	anti-ebov	human-adapted
nsp14-exon	imminent	infection-induced	investigational	corona-and

COV	Disease	Infection	Antimalarial	Virus
sars-related	multi-factorial	Diagnosing	hydroxychloroquine	pox
sars-like	afflict	Abstractstudies	nitazoxanide	lyssaviruses
coronaviral	ailment	Reinfections	anti-dengue	hhv
sars-cov	malarial	Causally	long-acting	constitution
sarsr-cov	abm	pre-infection	bioavailable	2003 - 2004
sars-and	self-limited	Predilection	broadspectrum	tilv
btcov	cryptosporidiosis	Ensue	anti-malarial	single-and
beta-coronavirus	melioidosis	Cant	anthelmintic	virus-2

Evaluation results

The system was implemented using Protégé and Python 3. The following metrics were used to describe the efficiency of the model. The COVOC model was built as a classifier that runs using information extracted from the coronavirus literature. In the information extraction phase, the latent topics or concepts present in the literature were retrieved. For experimental evaluation, we took a sample of the dataset obtained from the extracted concepts. In order to prove the efficiency of the model, we considered 15 classes, Cell, Influenza, Disease, Virus, Protein, Infection, Respiratory, RNA, Viral, Medicine, COV, Patient, Antimalarial, patient id ,pathogen and human as our topic attributes, and for each attribute, roughly 400 instances were considered. They were verified for various metrics, such as Accuracy, Precision, Recall and F1 scores.

Accuracy

Accuracy is a measure of the validity of a model, denoted by TP-True Positives, TN-True Negatives, FP-False Positives, and FN-False Negatives. It is the ratio of correct predictions to total number of predictions, as given in Eq. (8):

$$\text{Accuracy} = TP + TN/TP + FP + FN + TN \quad (8)$$

Generally, a model with higher accuracy is a better one; however, real accuracy is obtained when both FP and FN are equal in number, which is rare. Hence, we considered other performance measures, such as precision, recall and F1 scores. The accuracy obtained for COVOC was compared with other ontology based text classification baseline models such as OBTC [20], OBJHA [XX1], OBOHS [22] and OBTM [23].

Precision

The Positive Prediction Value (PPV), or precision, denotes the ratio between positively predicted values to total positive predicted values. It is calculated using the equation given in Eq. (9) [24]:

$$\text{Precision} = TP/TP + FP \quad (9)$$

Recall

Recall is given many other names, such as True Positive Rate (TPR), Sensitivity, and hit rate. It is the ratio of correctly classified positive samples to the total number of positive samples in the classification. The method to calculate recall is denoted by Eq. (10) [24]:

$$\text{Recall} = TP/TP + FN \quad (10)$$

F1 score

This is also called F-measure, and is the harmonic mean of the above two metrics. The higher values for F1 indicate model effectiveness. This reflects the class distribution of the sample set efficiently, which is the basic requirement for problems that involve unknown class distributions [24].

$$\text{F1 score} = 2*(\text{Recall}\times\text{Precision})/(\text{Recall} + \text{Precision}) \quad (11)$$

Multi-class problems were verified for their performance using these measures. The different assessment methods and their results are provided in **Table 3**. From **Table 3**, it was found that our proposed COVOC

model achieved a performance improvement of 85 to 90 %, which is 7 to 8 % more, compared to the baseline models.

Table 3 Classification metrics obtained for COVOC model compared with other baselines.

Data Model	Experiment	Values
OBTC	Accuracy	75.08589
	Precision	50.41122
	Recall	58.08895
	F1 Score	54.25008
OBJHA	Accuracy	71.72935
	Precision	59.20543
	Recall	54.33179
	F1 Score	56.76861
OBOHS	Accuracy	87.50445
	Precision	70.4808
	Recall	68.6908
	F1 Score	69.5858
OBTM	Accuracy	64.34622
	Precision	69.93915
	Recall	64.12481
	F1 Score	67.03198
COVOC	Accuracy	92.72642
	Precision	86.91442
	Recall	79.21784
	F1 Score	83.06613

Confusion matrix

This is also known as a contingency table [25]. The correctly classified samples are shown in the diagonal. The number of instances taken for testing is given at right vertical axis. The total number of classes is given in the x and y axes, respectively. Samples other than those occupying diagonal are misclassified samples. **Figure 4** denotes the Confusion Matrix plotted for **Table 2**, which is the final output of our model. The obtained dataset had more than 1,300 classes; hence, we took only a sample to test our model. All these classes were some of the key topics present in the coronavirus literature. The knowledge obtained from our model can be further used to solve other problems in text classification related to the disease.

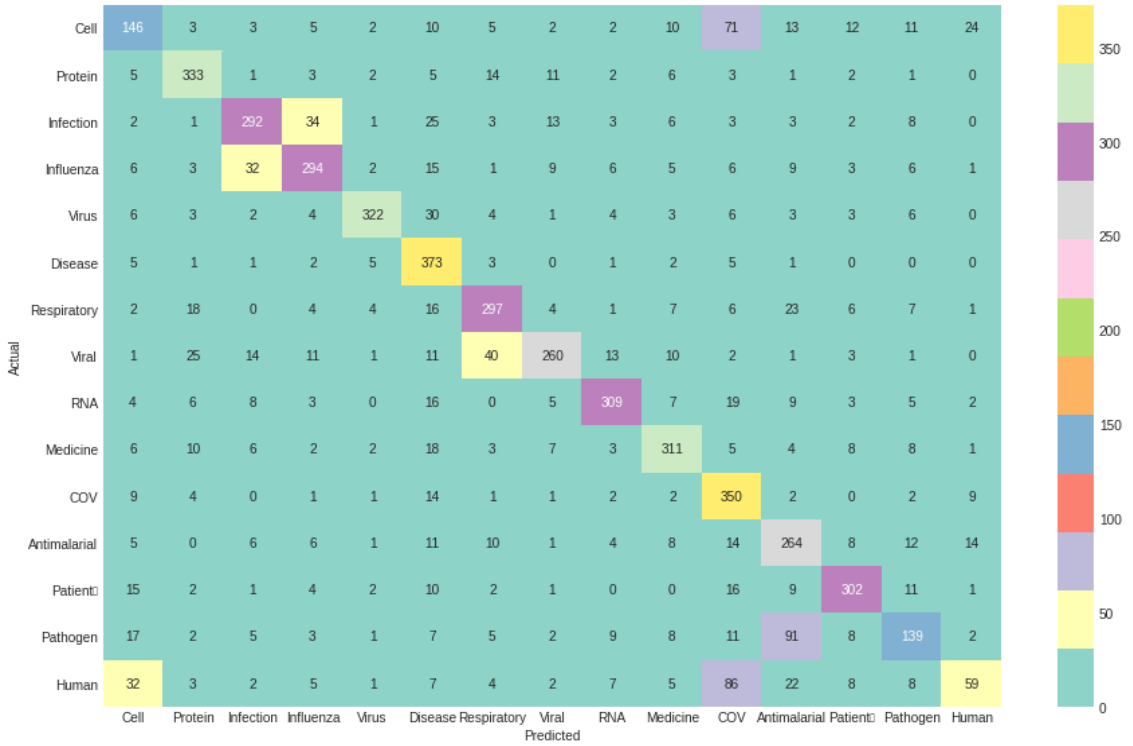

Figure 4 Confusion matrix of COVOC model.

PR curve

The use of the Precision Recall curve mainly concerned the information extraction ability of our proposed COVOC system. In our model, the classes sometimes became unbalanced, due to the presence of varying disease information in literature. The PR curve helped to prove that the unbalance in classes did not affect the overall efficiency of the model. The PR curve given in **Figure 5** shows that the trade-off between precision and recall was healthy.

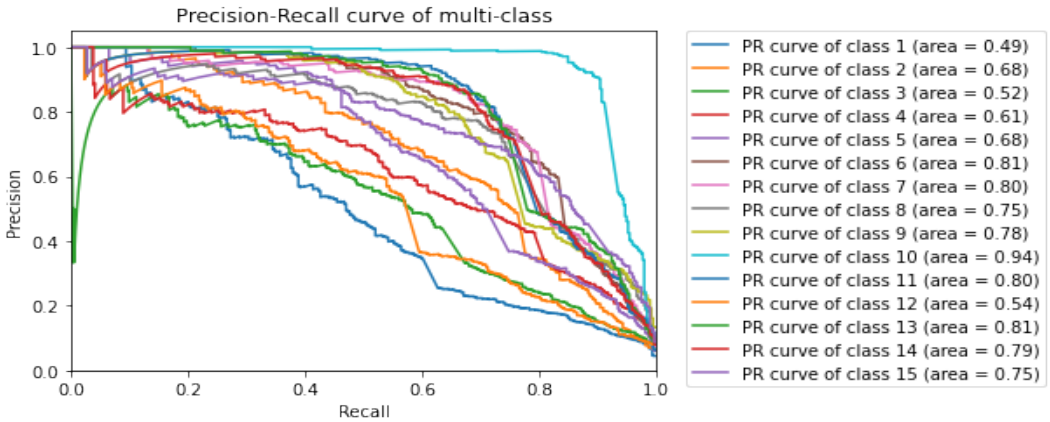

Figure 5 PR curve of COVOC model.

ROC curve

This is the Receiver Operating Characteristic curve, which states the performance of a model against a threshold in a 2-dimensional plot. It compares the TPR and False Positive Rate. The probability models have inbuilt threshold decided by the algorithms; hence, when the classes reach above the threshold, it is classified as positive. In medical diagnosis, it is important to assess the benefits and costs

before making decisions, ROC is most suited to such problems. For our problem, the threshold was kept constant and the curve was plotted. From **Figure 6**, it is seen that the lower recall rate proved that the cut-off value was higher on the positive class, and that most of the classification happened in the optimistic zone. Thus, the COVOC model exhibited a good trade-off between True and False positive values.

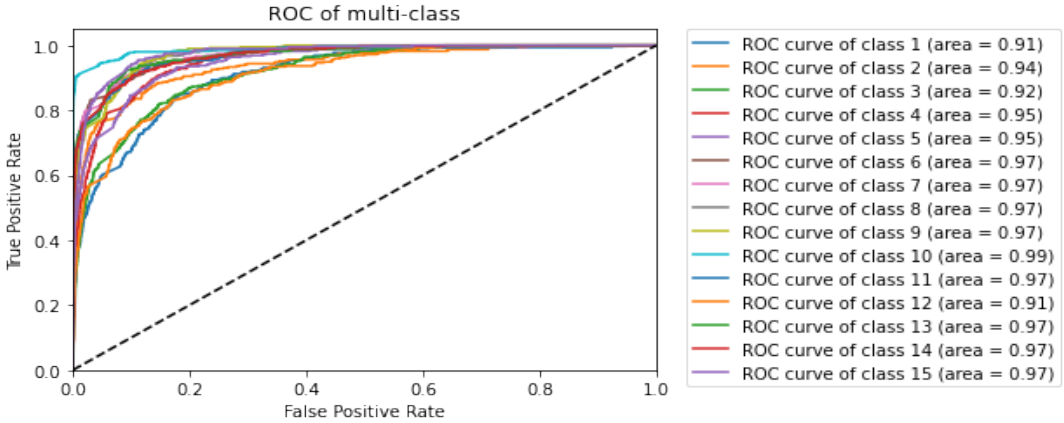

Figure 6 ROC of COVOC model.

Discussion

The basic idea behind the formulation of our work is information extraction from documents. However, due to the vastness of the collection, and the highly complex nature of the literature, the system calls for sophisticated data models. Even in domain knowledge representations, issues such as Synonymy, Polysemy, and Multi-word concepts arise. Some models suffer from extensive search space and the inability to reduce feature set due to their disregard for semantic matching of concepts. To solve these, the proposed model was developed as an ontology based text classifier. This kind of model can help to disambiguate complex medical texts. The concepts are annotated through automated semantic techniques with minimal human intervention, which helps to preserve the contextual meaning of the texts, crucial to accurate disease diagnosis. A drastic misclassification may lead to loss of life in the case of medical diagnosis. To measure the model's accuracy from an information retrieval perspective, the accuracy of the model was obtained from the contingency table. This is the process of treating an information retrieval problem as a 2 class classifier problem, where the most relevant articles retrieved by the model are considered. However, in most information retrieval problems, the data will be highly skewed, which can be solved using automated labeling techniques, as proposed in this model, which help to minimize false positives. Hence, the proposed model was treated as a machine learning classifier, where the procedure is the same for evaluating the model, but without major bias. The multi-class classifier proposed using the COVOC model showed remarkable performance in almost all of the metrics evaluated. The model was able to effectively extract most of the important concepts related to the nature of disease, medications, and virus types that are spread all over the literature collection. This can be seen from **Table 2**. The kind of retrieved information also helped to achieve the labeling of unstructured medical concepts, which is tedious in conventional data labeling techniques, given the complex nature of short expressions and other technical terms in the literature. The idea of removing non-medicine related words during the pre-processing phase using Metathesauruses has shown excellent improvement in removing bias in classification. The rich domain information obtained from the CORD dataset facilitated key semantic concept extraction related to coronavirus, about which only scarce information is available. The concepts were analyzed with genuine correlations to derive more meaningful conclusions regarding the disease, due to its semantic nature. The zero frequency issue that existed regarding the key concepts of the disease has been eliminated to some extent with the final connected semantic dataset obtained. The resulting dataset can be used for further research related to the coronavirus domain, such as correlation extraction, classification, named entity extraction, and document classification problems.

Conclusions

Understanding the nature of this disease, when there is no available cure, is vital to encourage accurate clinical diagnosis and drug discovery prospects. From the experimental analysis, it was proved that the COVOC model was able to capture the morphology, semantics, syntax, and pragmatics of the given literature, in order to extract key concepts related to the disease; also, zero frequency was solved using *improved*-LDA based topic modeling and subsequent ontological representation. The model was able to preserve the contextual information in the document space, while also providing holistic feature representation facilities. The use of M-SVM and *improved*-LDA for multi class text classification proved to have outstanding performance through the results of performance metrics such as Precision, Recall, F1 scores, and ROC. The performance improvement was in the range of 85 to 90 %, which is 7 to 8 % improvement compared to the baseline models. However, rigorous evaluation of the model by the medical practitioners is yet to be tried. The resulting dataset can be used for clinical diagnosis and further research on the disease. In future, research articles related to other diseases could be analyzed using this model and the creation of a more pragmatic knowledge base could be explored with other topic modeling algorithms.

References

[1] M Abdollahi, X Gao, Y Mei, S Ghosh and J Li. An ontology-based two-stage approach to medical text classification with feature selection by particle swarm optimisation. *In*: Proceedings of the IEEE Congress on Evolutionary Computation, Wellington, New Zealand. 2019, p. 119-26.

[2] F Ali, D Kwak, P Khan, S El-Sappagh, A Ali, S Ullah, KH Kim and KS Kwak. Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowl. Based Syst*. 2019; **174**, 27-42.

[3] M Allahyari, KJ Kochut and M Janik. Ontology-based text classification into dynamically defined topics. *In*: Proceedings of the IEEE International Conference on Semantic Computing, Newport Beach, CA, USA. 2014, p. 273-8.

[4] F Camous, S Blott and AF Smeaton. *Ontology-based MEDLINE document classification*. *In*: S Hochreiter and R Wagner (Eds.). Bioinformatics research and development. Springer, Berlin Heidelberg, 2007, p. 439-52.

[5] F Cardillo and U Straccia. Towards ontology-based explainable classification of rare events, Available at: https://hal.archives-ouvertes.fr/hal-02104520, accessed October 2020.

[6] CK Cheng, X Pan and F Kurfess. *Ontology-based semantic classification of unstructured documents*. *In*: A Nurnberger and M Detyniecki (Eds.). Adaptive multimedia retrieval. Springer, Berlin, Heidelberg, 2004, p. 120-31.

[7] J Fang, L Guo, X Wang and N Yang. Ontology-based automatic classification and ranking for web documents. *In*: Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery, Haikou, China. 2007, p. 627-31.

[8] M Gayathri and RJ Kannan. Ontology based concept extraction and classification of ayurvedic documents. *Proc. Comput. Sci.* 2020; **172**, 511-6.

[9] A Hawalah. Semantic ontology-based approach to enhance Arabic text classification. *Big Data Cogn. Comput*. 2019; **3**, 53.

[10] HH Kim and HY Rhee. An ontology-based labeling of influential topics using topic network analysis. *J. Inf. Process. Syst*. 2019; **15**, 1096-107.

[11] T Li and Z Chen. An ontology-based learning approach for automatically classifying security requirements. *J. Syst. Softw*. 2020; **165**, 110566.

[12] A Rashwan, O Ormandjieva and R Witte. Ontology-based classification of non-functional requirements in software specifications: A new corpus and SVM-based classifier. *In*: Proceedings of the IEEE 37th Annual Computer Software and Applications Conference, Kyoto, Japan. 2013, p. 381-6.

[13] K Sangounpao and P Muenchaisri. Ontology-based naive bayes short text classification method for a small dataset. *In*: Proceedings of the 20th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, Toyama, Japan. 2019, p. 53-8.

[14] N Shanavas, H Wang, Z Lin and G Hawe. Ontology-based enriched concept graphs for medical document classification. *Inf. Sci*. 2020; **525**, 172-81.

[15] M Thangaraj and M Sivakami. A comprehensive framework for ontology based classifier using unstructured data. *Int. J. Eng. Adv. Technol*. 2019; **9**, 6918-25.

[16] H Jelodar, Y Wang, C Yuan, X Feng, X Jiang, Y Li and L Zhao. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimed. Tools Appl*. 2019; **78**, 15169-211.
[17] VS Anoop, S Asharaf and P Deepak. Unsupervised concept hierarchy learning: A topic modeling guided approach. *Proc. Comput. Sci.* 2016; **89**, 386-94.
[18] RMA Mohammad. An enhanced multiclass support vector machine model and its application to classifying file systems affected by a digital crime. *J. King Saud Univ. Comput. Inf. Sci*. 2019, DOI: https://doi.org/10.1016/j.jksuci.2019.10.010.
[19] E Bernadó-Mansilla and JM Garrell-Guiu. Accuracy-based learning classifier systems: Models, analysis and applications to classification tasks. *Evol. Comput*. 2003; **11**, 209-38.
[20] M Allahyari, KJ Kochut and M Janik. Ontology-based text classification into dynamically defined topics. *In*: Proceedings of the IEEE International Conference on Semantic Computing, Newport Beach, CA, USA. 2014, p. 273-8.
[21] NW Chi, KY Lin and SH Hsieh. Using ontology-based text classification to assist Job Hazard Analysis. *Adv. Eng. Inf*. 2014; **28**, 381-94.
[22] N Sanchez-Pi, L Marti and ACB Garcia. Improving ontology-based text classification: An occupational health and security application. *J. Appl. Log*. 2016; **17**, 48-58.
[23] F Ali, D Kwak, P Khan, S El-Sappagh, A Ali, S Ullah, KH Kim and KS Kwak. Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowl. Based Syst*. 2019; **174**, 27-42.
[24] C Goutte and E Gaussier. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. *In*: Proceedings of the European Conference on Information Retrieval, Berlin, Heidelberg. 2005, p. 345-59.
[25] A Tharwat. Classification assessment methods. *Appl. Comput. Inf*. 2021; **17**, 168-92.
[26] B Dutta and M DeBellis. CODO: An ontology for collection and analysis of COVID-19 data. *In*: Proceedings of the 12th International Conference on Knowledge Engineering and Ontology Development, Budapest, Hungary. 2020.