Classification of a COVID-19 dataset by using labels created from clustering algorithms

Layth Rafea[1], Abdulrahman Ahmed[2], Wisam D. Abdullah[3]
[1,3]Cisco Networking Academy, Tikrit Universiti, Tikrit, Iraq
[2]Department of Network Engineering\Collage of Engineening, Al-Iraqia University, Iraq

Article Info

Article history:

Received Apr 11, 2020
Revised Jun 15, 2020
Accepted Jun 27, 2020

Keywords:

Classification
Clustering
COVID-19 pandemic
Dimensionality reduction
Pre-processing

ABSTRACT

Novel coronavirus (COVID-19) is a newly discovered infectious disease that has received much attention in the literature because of its rapid spread and daily global deaths attributable to such disease. The White House, together with a coalition of leading research groups, has published the freely available COVID-19 Open Research Dataset to help the global research community apply the recent advances in natural language processing and other AI techniques in generating novel insights that can support the ongoing fight against this disease. In this paper, the hierarchical and k-means clustering techniques are used to create a tool for identifying similar articles on COVID-19 and filtering them based on their titles. These articles are classified by applying three data mining techniques, namely, random forest (RF), decision tree (DT) and bagging. By using this tool, specialists can limit the number of articles they need to study and pre-process these articles via data framing, tokenisation, normalisation and term frequency-inverse document frequency. Given its 2D nature, the dimensionality of this dataset is reduced by applying t-SNE. The aforementioned data mining techniques are then cross validated to test the accuracy, precision and recall performance of the proposed tool. Results show that the proposed tool effectively extracts the keywords for each cluster, with RF, DT and bagging achieving optimal accuracies of 98.267%, 97.633% and 97.833%, respectively.

Corresponding Author:

Layth Rafea Hazim
Cisco Networking Academy, Tikrit University
Computer Center Building, Floor#2
Cisco Academy Center, Right Side, Tikrit, Iraq
Email: Layth.R.Hazim@tu.edu.iq

1. INTRODUCTION

The pneumonia outbreak in Wuhan, China late last year has eventually evolved into one of the worst pandemics in human history. This outbreak was triggered by the novel coronavirus termed COVID-19, which belongs to the Orthocoronavirinae subfamily and is distinct from the two coronaviruses reported in recent history, namely, the middle east respiratory syndrome and severe acute respiratory syndrome (SARS-CoV) coronaviruses as described in [1], The first case of COVID-19 was reported on 12 December 2019 and was later distinguished from SARS-CoV by the Chinese Center for Disease Control and Prevention (CDC). The Coronaviridae family comprises single, plus-stranded and large RNA viruses that are isolated from multiple species and trigger common cold and diarrhoea amongst humans [2, 3]. The COVID-19 outbreak was eventually declared a public health emergency by the World Health Organisation (WHO) on 8 March 2020 after 100,000 cases and 3,830 deaths were reported in more than 100 countries around the globe [4], thereby necessitating governments to implement drastic measures to control the disease whilst sacrificing their economic and social development. However, the transmission characteristics of this disease remain unknown to date [5], In addition, countries are struggling in controlling the spread of the disease given the accelerated rate of global urbanisation, their high population concentration and shortage of medical resources.

Although thousands of literature on viruses and their transmission, prevention and possible treatment have been published in recent years, most of these articles have only focused on SARS-CoV. As the number of infections and deaths attributable to COVID-19 continues to rise, specialists are racing in their search for a cure to this pandemic. However, given the large number of recent literature on viruses, these specialists spend much of their time in extracting articles that are actually related to COVID-19 and benefit from their findings [6], In this case, Text mining tools help biomedical researchers and clinicians to save time and effort which are devoted for acquiring valuable information from several documents. Activating, interpreting, and comprehensible information from many related sources of biomedical text are requiring tasks, which demand improving and creating automatic tools. [7], Articles in "CORD-19" are distributed and over many topics, and thousands of articles published a weekly basis. The "clustering" articles with the same topics will map commonalities and assist to the researchers to conduct new researched. Text mining by using "clustering", helps researchers use bibliographic datasets to get a rapid review of the topics [8], "Clustering" articles can decide what topics are covered in a good way and worth a review. Furthermore, these clusters dataset help researchers and decision-makers to specify related topics in research on COVID-19. As with the fundamental goal of the text classification, which is also known as text categorization, is the classification of texts of interest to correct classes [9], Consequently, text classification has gotten a great deal of interest in hierarchically organizing those articles. Up until now, text classifying was successfully implemented in different areas like topic detection [10], and document classifying. In classification step, a classifier performs the procedure of classification with the use of previously known labeled data, and documents are classified into appropriate classes [11], As the articles in CORD-19 are represented with numeric values, any classifier used in pattern recognition problems can be integrated to text/document classification process [12], However, selection of appropriate classifier increases success ratio of classification.

Multi-stage text mining presents is an effective method for classifying and clustering a large number of articles based on certain parameters. [13], Accordingly, in this paper, text mining is applied to analyse the COVID-19 Open Research Dataset (CORD-19), a dataset containing 44,000 scientific articles on viruses [14], and filter these articles according to their titles to extract only those works that are particularly relevant to specialists in their search for an effective treatment against COVID-19. In other words, this study primarily aims to reassemble scientific articles on viruses in a way that only those articles that are directly related to COVID-19 will be presented to specialists.

2. RELATED WORKS

Few published works have applied data mining and text mining in an attempt to find solutions to the COVID-2019 pandemic. This paper only focuses on those techniques that aim to establish relationships amongst the text and categorical attributes of the relevant literature. This section summarises the literature related to this work.

A hybrid classification framework based on clustering (HCFC) was proposed in [15], This framework initially applies a clustering algorithm to divide an entire training sample into k clusters. Afterwards, a clustering-based attribute selection measure called hybrid information gain ratio was constructed to train a C4.5 decision tree. Two versions of HCFC, namely, HCFC-K and HCFC-D, were then built and tested on 8 benchmark datasets related to healthcare and disease diagnosis and 15 datasets from other fields. Results show that both HCFC-K and HCFC-D either compare or outperform the other three hybrid and six single models considered in the study. Between these two algorithms, HCFC-D shows a better resistance to class noise. The authors in [16], tested the feasibility of using a modelling approach in identifying patient safety events (PSEs) related to HIT usability from the free text of safety reports and studied how such approach can be used by patient safety analysts in analysing event data. With a dataset containing 5,911 manually annotated reports, they identified PSEs related to HIT usability by using three feature representations, namely, bag of words (BOWs), topic modelling and document embedding. Together with patient safety analysts, the authors reviewed the results of their approach and gathered feedback on its usefulness and integration into workflows. Combining term frequency-inverse document frequency (TD-IDF) BOWs with document-embedding features that are modelled via support vector machine (SVM) with radial basis function (RBF) yielded the best precision-recall performance with an under the curve (AUC) and f1 score of 72% and 66%, respectively. Meanwhile, compared with the SVM RBF model, the application of document-embedding features resulted in an AUC and f1 score of 70% and 66%, respectively. Both approaches favoured sensitivity and specificity over precision. Patient safety analysts also reported the usefulness of such approach in their point of report entry, visual dashboard layers and data retrieval. In sum, document embedding and text mining approaches can support the identification of PSEs related to HIT usability. An econometric model was proposed in [17], to predict the spread of COVID-19. The authors applied an autoregressive integrated moving average model on an epidemiological dataset from Johns Hopkins Center covering the dates 20 January to 10 February 2020 to predict the prevalence and incidence trends of the disease. The overall prevalence of COVID-19 demonstrates an increasing trend that reaches epidemic proportions, and the difference between the cases reported in one day and in the previous day D (Xn-Xn-1) does not indicate any constant increase in the number of confirmed cases. A descriptive analysis was performed to check for any potential bias and to evaluate the incidence of new confirmed cases of COVID-19. The authors in [18], compiled and analysed epidemiological outbreak information on COVID-19 by using the open datasets of Johns Hopkins University, WHO, CDC, National Health Commission and DXY. The number of confirmed cases, deaths and recoveries was also investigated via an exploratory data analysis accompanied with data visualisations.

3. RESEARCH METHOD

Figure 1 presents the main stages of the methodology adopted in this work. Recall, precision and accuracy are used as performance measures. The results are processed by using Python (Notebook) in order to build document classification models (including pre-processing, features extraction, features selection and features modification models) based on three classification algorithms, namely, random forest (RF), decision tree (DT) and bagging.

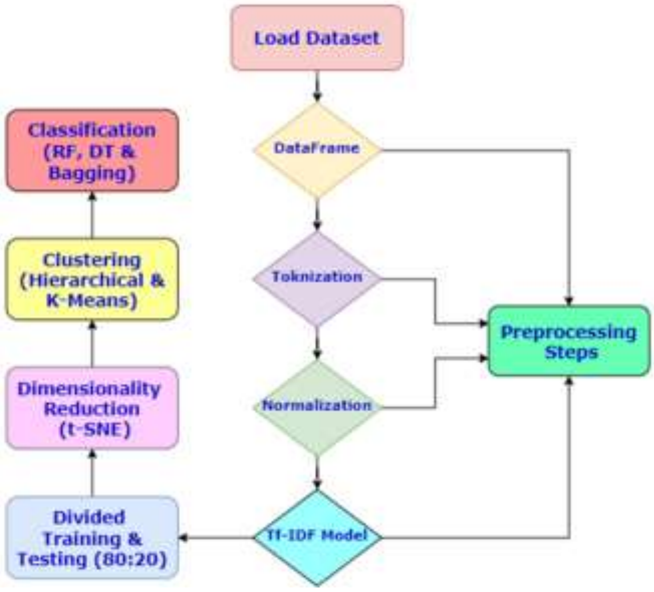

Load Dataset
DataFrame
Toknization
Normalization
TF-IDF Model
Preprocessing Steps
Classification (RF, DT & Bagging)
Clustering (Hierarchical & K-Means)
Dimensionality Reduction (t-SNE)
Divided Training & Testing (80:20)

Figure 1. Stages of the research methodology

3.1. Pre-processing

The dataset was initially pre-processed to present the text articles in a clear word format. Specifically, the contents of these articles were pre-processed to remove irrelevant words in order to facilitate the classification and to reduce their dimensionality [19], The pre-processing stage involves four steps, namely, data framing, tokenisation, normalisation and TF-IDF stop words feature selection.

3.1.1. Data frame

A data frame is a 2D array-like table where each column contains the values for a single variable and each row contains a set of values for each column [20], A data frame should possess the following characteristics:

a) The names of each column should not be empty.
b) Each row should be given a unique name.
c) The data can be of numeric, character or factor type.
d) All columns should contain the same number of items.

3.1.2. Tokenisation

In the tokenisation step, the text data are split into tokens or simple independent units of words or terms depending on their distance in order to separate the words in the text from one another [13], These tokens are also crucial in natural language processing (NLP).

3.1.3. Normalisation

The next step in NLP is pre-processing the articles to normalise the data. In this step, different forms of the same letter are normalised by converting all characters into lower or upper case and deleting all symbols and numbers [20].

3.1.4. TF-IDF model

TF-IDF generates weighted term vectors that will be subsequently used for clustering and classification. The TF-IDF model is widely applied in feature extraction and selection [21], In this step, the articles are transformed into a document vector, and each article is represented by an array of weights. In this case, the collection of text documents can be represented in matrix form where each document is assigned to a single row and each feature in a list of vocabulary is assigned to each column. Each feature is also associated with a weight to indicate its relative importance in the entire document. The parameters in the TF-IDF model employed in this work include stop word (='English') and max_features (=2**12) to guarantee high clustering and classification accuracy.

3.2. Training and testing datasets

The training and testing sets are separated with cut-offs of 80% and 20%. These sets are later used as inputs for dimensionality reduction and for both the classification and clustering algorithms. Each word is treated as an atomic unit, and each term is assigned a weight according to the TF-IDF model.

3.3. Dimensionality reduction with t-SNE

Dimensionality reduction is an optional step that can be achieved by using classification and clustering models. To guarantee an excellent performance, previous studies have applied dimensionality reduction to address time and memory complexities. Moreover, conducting pre-processing via dimensionality reduction is also more efficient than developing inexpensive classifiers. T-distributed stochastic neighbour embedding (t-SNE) is a nonlinear dimensionality reduction method for embedding high-dimensional data that is widely utilised for visualisation in a low-dimensional feature space [22], Moreover, using t-SNE can reduce a high-dimensional features vector into a 2D plane. During this process, t-SNE keeps similar instances together whilst pushing different instances far from one another. As shown in Figure 2, the resulting 2D plane reveals which articles are clustered near one another. In this study, t-SNE is applied with the parameters verbose (=1) and perplexity (=5) to guarantee an excellent performance.

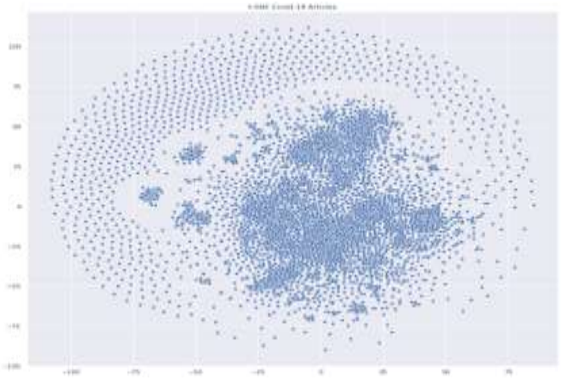


Figure 2. Dimensionality reduction via t-SNE

3.4. Clustering algorithms

The clustering algorithm aims to maximise intra-cluster similarity whilst minimising intercluster similarity, both of which are calculated by using several measures, including the Euclidean distances between the data points and nearest cluster centroids [13], This step can also be used to create a tool for identifying articles that are similar to the target article, thereby limiting the number of articles that need to be analysed. Two clustering algorithms are applied in this work, namely, hierarchical clustering and k-means clustering.

3.4.1. Hierarchical clustering

Hierarchical methods can be either agglomerative (bottom-up approach) or divisive (top-down approach) [23], Hierarchical clustering, also known as hierarchical cluster analysis, groups similar objects into groups called clusters that are distinct from one another [13]. Agglomerative clustering treats each document as an independent cluster and merges all homogeneous clusters until no further merging can be performed. To guarantee excellent performance, several parameters are employed in this work, namely, n_clusters (=10), affinity (='euclidean') and linkage (='ward').

3.4.2. K-Means clustering

K-means clustering algorithm clusters documents based on the entropy global term weighting method and aims to determine how many k clusters are available in the data. This algorithm iteratively moves k centres and selects data points closest to the centroid [24], In this work, an unsupervised k-means algorithm is employed in conjunction with input TF-IDF vectors to define the auxiliary sub-task. An auxiliary sub-task definition is then formulated for each of the main NLU tasks [20]. Meanwhile, the k-means unsupervised technique is applied to obtain the labels necessary for establishing supervised classification techniques given that the dataset lacks any label. The parameters clusters (=10), n_jobs (=4) and verbose (=10) are employed in this work.

3.5. Classification algorithms

Text classification or categorisation uses a set of class-labelled documents from a specific domain to build a model that provides class predictions for arbitrary documents from the same domain [25], This study employs three classification algorithms, namely, RF, DT and bagging.

3.5.1. Random forest

RF is a popular ensemble modelling algorithm that achieves excellent predictive performance by combining multiple models from the same domain [26], An RF is represented by a set of unpruned DTs that are grown based on multiple bootstrap samples that are drawn (with replacements) from the training set via randomised split selection. RF is a rapid and accurate technique employed for document categorisation and text classification. RF can train text data sets much faster than other techniques, including deep learning, yet is slow in making predictions after training [22], This study employs the parameters n_estimators (=100), random_state (=42) ands n_jobs (=4) to achieve optimal accuracy.

3.5.2. Decision tree

DT has been successfully employed in many fields for classification. This technique has a structure that resembles a hierarchical decomposition of the data space [27], DT performs inductive learning from the data and obtains a tree-like structure that is equivalent to a set of decision rules [28], The main idea of this algorithm is to create a tree based on the attributes of the categorised data points. However, which of these attributes should be ascribed to the parent or child level needs to be determined. DT is a very fast algorithm for both learning and prediction but is extremely sensitive to small perturbations in the data and is prone to overfitting. Nevertheless, such challenges can be addressed by employing validation and pruning. In this study, the DT algorithm is combined with different parameters.

3.5.3. Bagging

Bagging is employed to classify documents and text datasets and is generated by using different bootstrap samples, with each bootstrap generating a uniform sample from the training set [22]. Bagging creates a sequence of classifiers in consideration of the modifications applied to the training set. This method has also been applied to improve the efficiency of standard machine learning algorithms. The classifiers created by bagging are combined into a compound classifier, which prediction is assigned a weighted combination of individual classifier predictions [29], Regardless of these advances, bagging has several disadvantages, including its computational complexity and loss of interpretability, which prevents this algorithm from recognising the importance of each feature. The parameters n_estimators (=100) and random_state (=1) are employed in this work.

4. RESULTS AND DISCUSSION

This section presents the dataset, learning algorithms, a performance evaluation criteria and evaluation measures used in this work. The clustering and classification algorithms that are applied in our work, depending on the title in the articles. Python (Notebook) is used to process the results.

4.1. Dataset description

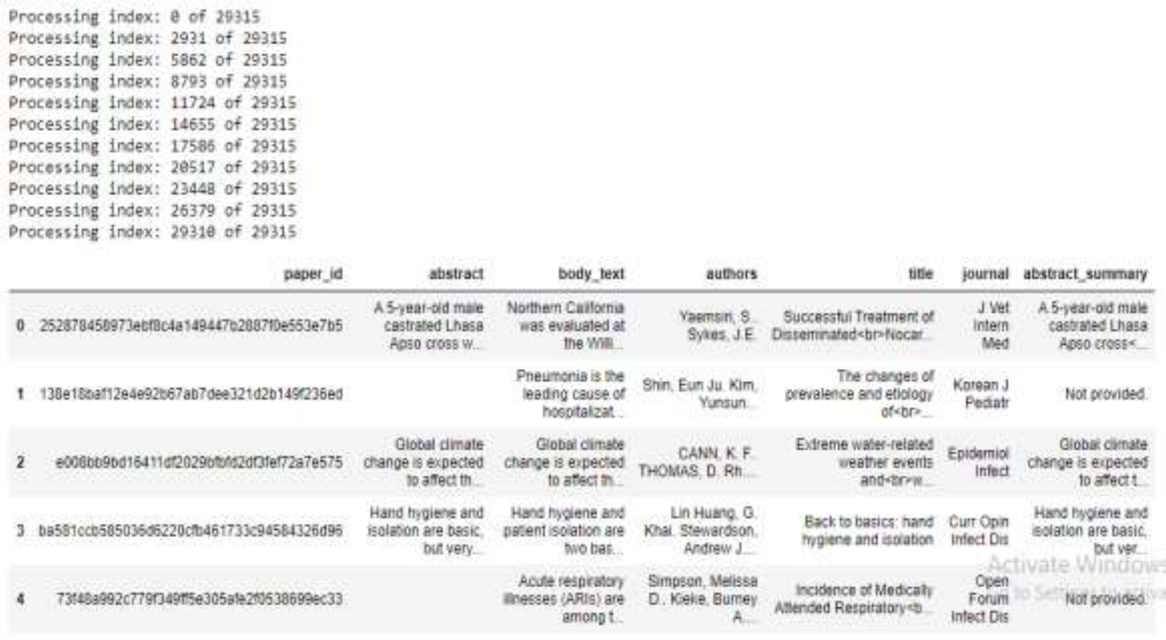
In collaboration with leading research groups around the world, the Allen Institute for AI has prepared the CORD-19 dataset, a free resource that consists of over 44,000 scholarly articles on COVID-19 and the coronavirus family, of which more than 29,000 articles are presented in full text with 15 features as shown in Table 1 [14], This dataset aims to help researchers apply the recent advances in NLP to obtain novel insights that can support the race for a solution to the COVID-19 pandemic.

Table1. Description of dataset

Features Name	Description	Records
Sha	The paper records have PDFs or include multiple files (some PMC files have multiple associated PDFs).	28462 non-null
Source_x	The articles source such as (Elsevier).	44220 non-null
Title	All articles title in this dataset (our work on this feature).	43996 non-null
Doi	Populated for all "BioRxiv/MedRxiv" paper records and most of the other records.	40750 non-null
Pmcid	Populated for all PMC paper records.	23319 non-null
Pubmed_id	ID of the articles populated for some of the records.	22943 non-null
License	Custom_license of articles.	44220 non-null
Abstract	The abstract of all articles that are already there.	35806 non-null
Publish_time	Publishing time for articles.	34197 non-null
Authors	The names of authors in the articles.	41074 non-null
Journal	The names of journals publishing articles.	33173 non-null
Microsoft Academic Paper ID	Populated for some of the records.	964 non-null
WHO #Covidence	Populated for all CZI records and none of the other records.	1767 non-null
Has_full_text	Number of the "PDFs" were processed with full text.	44220 non-null
Full_text_file	The signal the "tar.gz" file in which the full text "json" resides.	32829 non-null

4.2. Pre-processing results

The dataset is pre-processed to facilitate data framing, remove the irrelevant words and symbols in the selected articles, revert some words back to their original forms and apply the TF-IDF model. The employed dataset contains 29,315 full-text articles. The data are entered in four stages, with each stage aiming to process and improve the presentation of these data. Figure 3 shows how the articles are entered into an easy-to-use data frame, whereas Figure 4 shows the dataset after the tokenisation and normalisation of the text abstracts, body texts and titles. In Figure 4, when compared to Figure 3, we notice the disappearance of the symbols and tokens in the tokinasation, as well as the transformation of all texts in the abstracts, body texts and titles into lower case.

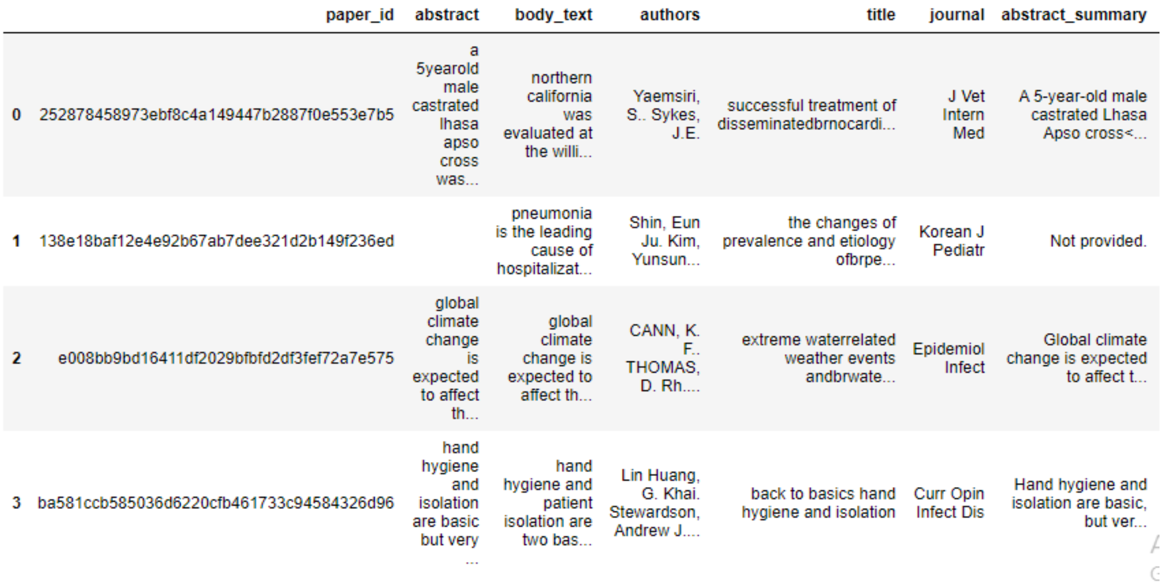


```
Processing index: 0 of 29315
Processing index: 2931 of 29315
Processing index: 5862 of 29315
Processing index: 8793 of 29315
Processing index: 11724 of 29315
Processing index: 14655 of 29315
Processing index: 17586 of 29315
Processing index: 20517 of 29315
Processing index: 23448 of 29315
Processing index: 26379 of 29315
Processing index: 29310 of 29315
```

	paper_id	abstract	body_text	authors	title	journal	abstract_summary
0	252878458973ebf8c4a149447b2887f0e553e7b5	A 5-year-old male castrated Lhasa Apso cross w...	Northern California was evaluated at the Willi...	Yaemsiri, S., Sykes, J.E.	Successful Treatment of Disseminated Nocar...	J Vet Intern Med	A 5-year-old male castrated Lhasa Apso cross<...
1	138e18baf12e4e92b67ab7dee321d2b149f236ed		Pneumonia is the leading cause of hospitalizat...	Shin, Eun Ju, Kim, Yunsun...	The changes of prevalence and etiology of ...	Korean J Pediatr	Not provided.
2	e008bb9bd16411df2029bfbfd2df3fef72a7e575	Global climate change is expected to affect th...	Global climate change is expected to affect th...	CANN, K. F., THOMAS, D. Rh...	Extreme water-related weather events and w...	Epidemiol Infect	Global climate change is expected to affect t...
3	ba581ccb585036d6220cfb461733c94584326d96	Hand hygiene and isolation are basic, but very...	Hand hygiene and patient isolation are two bas...	Lin Huang, G. Khai, Stewardson, Andrew J....	Back to basics: hand hygiene and isolation	Curr Opin Infect Dis	Hand hygiene and isolation are basic, but ver...
4	73f48a992c779f349ff5e305afe2f0538699ec33		Acute respiratory illnesses (ARIs) are among t...	Simpson, Melissa D., Kieke, Burney A....	Incidence of Medically Attended Respiratory<b...	Open Forum Infect Dis	Not provided.

Activate Windows
Go to Settings to activa

Figure 3. Dataset to data frame

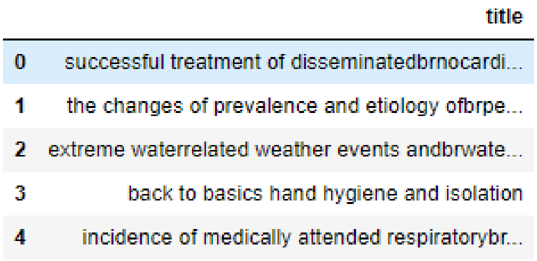


	paper_id	abstract	body_text	authors	title	journal	abstract_summary
0	252878458973ebf8c4a149447b2887f0e553e7b5	a 5yearold male castrated lhasa apso cross was...	northern california was evaluated at the willi...	Yaemsiri, S.. Sykes, J.E.	successful treatment of disseminatedbrnocardi...	J Vet Intern Med	A 5-year-old male castrated Lhasa Apso cross<...
1	138e18baf12e4e92b67ab7dee321d2b149f236ed		pneumonia is the leading cause of hospitalizat...	Shin, Eun Ju. Kim, Yunsun...	the changes of prevalence and etiology ofbrpe...	Korean J Pediatr	Not provided.
2	e008bb9bd16411df2029bfbfd2df3fef72a7e575	global climate change is expected to affect th...	global climate change is expected to affect th...	CANN, K. F.. THOMAS, D. Rh....	extreme waterrelated weather events andbrwate...	Epidemiol Infect	Global climate change is expected to affect t...
3	ba581ccb585036d6220cfb461733c94584326d96	hand hygiene and isolation are basic but very ...	hand hygiene and patient isolation are two bas...	Lin Huang, G. Khai. Stewardson, Andrew J....	back to basics hand hygiene and isolation	Curr Opin Infect Dis	Hand hygiene and isolation are basic, but ver...

Figure 4. Tokenisation and normalisation

4.3. Dimensionality reduction by t-SNE

A scatter plot is generated by using the plain text from the titles of each article as shown in Figure 5. The sklearn feature in Tf-IDF Vectorizer is then used to transform each instance into a features vector, and the dataset is divided afterwards for training (80%) and testing (20%). Dimensionality reduction is eventually applied to the feature vectors by using t-SNE as shown in Figure 2. In this part, a solution to the problem was drawn up based on the titles in the articles and then classified accordingly, as shown in Figure 5.



	title
0	successful treatment of disseminatedbrnocardi...
1	the changes of prevalence and etiology ofbrpe...
2	extreme waterrelated weather events andbrwate...
3	back to basics hand hygiene and isolation
4	incidence of medically attended respiratorybr...

Figure 5. Titles of full-text articles

4.4. Clustering results

The clustering performance of hierarchical and k-means clustering on an unlabelled CORD-19 dataset is initially evaluated. Hierarchical clustering recursively merges a pair of clusters that minimally increases a given linkage distance. In this work, a total of 10 clusters are employed, and the 'euclidean' distance and linkage ('ward') are used to compare the performance of hierarchical clustering with that of k-means clustering as shown in Figures 6 and 7. K-means clustering is employed to generate the labels for the supervised classification models. The parameters k, n_jobs and verbose are set to 10, 4 and 10, respectively. Each cluster is also subjected to topic modelling to obtain the keywords for each cluster.

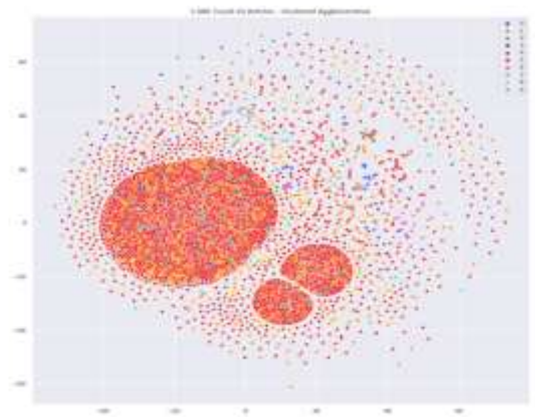


Figure 6. Hierarchical clustering

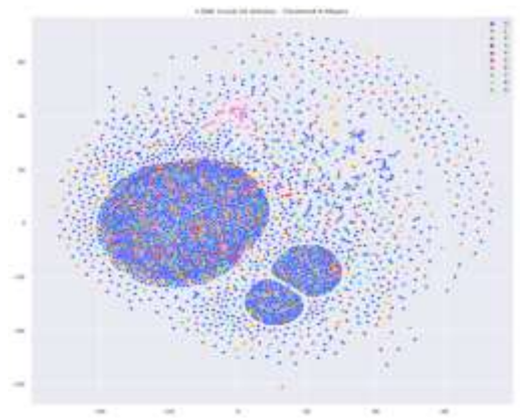


Figure 7. K-means clustering

4.5. Classification results

The classification evaluation performance of RF, DT and bagging on CORD-19 is then evaluated. The fit function in Python is used with accuracy, precision and recall as performance measures. Table 2 presents the classification results for the training dataset. In line with our expectations, RF shows the best performance amongst the three classification algorithms, followed by bagging and DT. Table 3 presents the classification performance of these algorithms for the testing dataset. In line with Table 2, Table 3 shows that RF achieves the best performance amongst the three classification algorithms employed when we chose classification based on the title in the articles in this work.

Table 2. Classification results for the training dataset

Models	RF	DT	Bagging
Accuracy	98.625	97.750	98.075
Precision	97.685	95.702	96.303
Recall	97.471	95.884	96.140

Table 3. Classification results for the testing dataset

Models	RF	DT	Bagging
Accuracy	98.267	97.633	97.833
Precision	96.612	95.332	95.766
Recall	97.111	95.380	95.803

5. CONCLUSION

This work aims to aid specialists in understanding the spread of COVID-19 around the world and in devising an effective treatment against the virus. Confirmed cases and deaths are mostly distributed in economically developed countries with strong medical testing capabilities. COVID-19 is known for its strong infectivity, long incubation period and difficulty of detection, all of which contribute to its rapid spread. This study attempts to understand the responses of developing countries to the pandemic by examining published articles that are directly related to COVID-19 in hopes of determining those factors that affect their ability to respond to such pandemic. The CORD-19 dataset, which is the most comprehensive dataset of machine-readable literature on COVID-19 and coronaviruses available to date, presents an unsupervised learning problem that can only be solved by using text and data mining tools [14]. This study attempts to address such problem by using a four-stage methodology. Firstly, the articles included in the CORD-19 dataset are pre-processed, in which these articles are cleaned and filtered via tokenisation, visualisation and application of TF-IDF, resulting in a final dataset containing 29,315 full-text articles that is divided into a training and testing dataset at a 80:20 ratio. Secondly, the dimensionality of the dataset is reduced by using t-SNET, hierarchical and k-means clustering are applied, where the articles in the final dataset are grouped into clusters based on the similarities in their titles, Labels are also generated by k-means clustering. Fourthly, the three data mining techniques RF, DT and bagging are applied to classify the similar articles based on their titles and to determine which of these algorithms show the best performance. The proposed topic model ACCM outperforms the other models that have been applied on the CORD-19 dataset in terms of (title) clustering and article classification, and the RF technique is the best results in the classification.

ACKNOWLEDGEMENTS

This work was funded by the Allen Institute for AI, which prepared the CORD-19 dataset in partnership with leading research groups, and Kaggle, which hosted the COVID-19 Open Research Dataset Challenge. We hope that the findings of this work can contribute to the extant knowledge on the COVID-19 pandemic.

REFERENCES

[1] N. Zhu *et al.*, "A novel coronavirus from patients with pneumonia in China, 2019," *N. Engl. J. Med.*, vol. 382, no. 8, pp. 727–733, 2020, doi: 10.1056/NEJMoa2001017.

[2] C. Drosten *et al.*, "Identification of a novel coronavirus in patients with severe acute respiratory syndrome," *N. Engl. J. Med.*, vol. 348, no. 20, pp. 1967–1976, 2003, doi: 10.1056/NEJMoa030747.

[3] Y. Chen, Q. Liu, and D. Guo, "Emerging coronaviruses: Genome structure, replication, and pathogenesis," *J. Med. Virol.*, vol. 92, no. 4, pp. 418–423, 2020, doi: 10.1002/jmv.25681.

[4] World Health Organization (WHO), "Novel Coronavirus (2019-nCoV) Situation Report - 1 21 January 2020," *WHO Bull.*, no. JANUARY, pp. 1–7, 2020.

[5] "CDC Centers for Disease Control and Prevention, 2020. Healthcare Professionals Frequently Asked Questions and Answers," *2020.*

[6] A. J. Rodriguez-Morales *et al.*, "Clinical, laboratory and imaging features of COVID-19: A systematic review and meta-analysis," *Travel Med. Infect. Dis.*, no. February, p. 101623, 2020, doi: 10.1016/j.tmaid.2020.101623.

[7] R. Mishra *et al.*, "Text summarization in the biomedical domain: A systematic review of recent research," *J. Biomed. Inform.*, vol. 52, pp. 457–467, 2014, doi: 10.1016/j.jbi.2014.06.009.

[8] L. Amador Penichet, D. Magdaleno Guevara, and M. M. García Lorenzo, "New similarity function for scientific articles clustering based on the bibliographic references," *Comput. y Sist.*, vol. 22, no. 1, pp. 93–102, 2018, doi: 10.13053/CyS-22-1-2763.

[9] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," *Knowledge-Based Syst.*, vol. 36, pp. 226–235, 2012, doi: 10.1016/j.knosys.2012.06.005.

[10] M. Ghiassi, M. Olschimke, B. Moon, and P. Arnaudo, "Automated text classification using a dynamic artificial neural network model," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 10967–10976, 2012, doi: 10.1016/j.eswa.2012.03.027.

[11] J. Han, M. Kamber, and J. Pei, "Data Mining Techniques, Third Edition," p. 847, 2011.

[12] D. Braha, *Data Mining for Design*. 2002.

[13] A. K. Abasi, A. T. Khader, M. A. Al-Betar, S. Naim, S. N. Makhadmeh, and Z. A. A. Alyasseri, "Link-based multi-verse optimizer for text documents clustering," *Appl. Soft Comput. J.*, vol. 87, p. 106002, 2020, doi: 10.1016/j.asoc.2019.106002.

[14] A. I. for A. in partnership with the C. Z. Initiative, "COVID-19 Open Research Dataset Challenge (CORD-19)," *semantic scholar,* 2020. [Online]. Available: https://pages.semanticscholar.org/coronavirus-research. [Accessed: 20-Mar-2020].

[15] J. Xiao, Y. Tian, L. Xie, X. Jiang, and J. Huang, "A Hybrid Classification Framework Based on Clustering," *IEEE Trans. Ind. Informatics*, vol. 16, no. 4, pp. 2177–2188, 2020, doi: 10.1109/TII.2019.2933675.

[16] A. Fong, T. Komolafe, K. T. Adams, A. Cohen, J. L. Howe, and R. M. Ratwani, "Exploration and Initial Development of Text Classification Models to Identify Health Information Technology Usability-Related Patient Safety Event Reports," *Appl. Clin. Inform.*, vol. 10, no. 3, pp. 521–527, 2019, doi: 10.1055/s-0039-1693427.

[17] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, and M. Ciccozzi, "Application of the ARIMA model on the COVID-2019 epidemic dataset," *Data Br.*, vol. 29, p. 105340, 2020, doi: 10.1016/j.dib.2020.105340.

[18] S. K. Dey, M. M. Rahman, U. R. Siddiqi, and A. Howlader, "Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach," *J. Med. Virol.*, no. February, pp. 1–7, 2020, doi: 10.1002/jmv.25743.

[19] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Inf. Process. Manag.*, vol. 50, no. 1, pp. 104–112, 2014, doi: 10.1016/j.ipm.2013.08.006.

[20] E. Métais, F. Meziane, S. Vadera, V. Sugumaran, and D. Hutchison, *Processing and*. 2019.

[21] B. Altınel and M. C. Ganiz, "Semantic text classification: A survey of past and recent advances," *Inf. Process. Manag.*, vol. 54, no. 6, pp. 1129–1153, 2018, doi: 10.1016/j.ipm.2018.08.001.

[22] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Inf.*, vol. 10, no. 4, pp. 1–68, 2019, doi: 10.3390/info10040150.

[23] M. A. Elaziz, N. Nabil, A. A. Ewees, and S. Lu, "Automatic Data Clustering based on Hybrid Atom Search Optimization and Sine-Cosine Algorithm," *2019 IEEE Congr. Evol. Comput. CEC 2019 - Proc.*, pp. 2315–2322, 2019, doi: 10.1109/CEC.2019.8790361.

[24] J. Rashid, S. M. A. Shah, and A. Irtaza, "An Efficient Topic Modeling Approach for Text Mining and Information Retrieval through K-means Clustering," *Mehran Univ. Res. J. Eng. Technol.*, vol. 39, no. 1, pp. 213–222, 2020, doi: 10.22581/muet1982.2001.20.

[25] F. Van Braam Houckgeest, M. J. Schultz, and P. E. Spronk, "Unsettled issues regarding intensive insulin therapy in the intensive care unit," *Netherlands J. Crit. Care*, vol. 13, no. 5, pp. 266–267, 2009.

[26] P. Cichosz, "A case study in text mining of discussion forum posts: Classification with bag of words and global vectors," *Int. J. Appl. Math. Comput. Sci.*, vol. 28, no. 4, pp. 787–801, 2018, doi: 10.2478/amcs-2018-0060.

[27] D. M. Magerman, "Statistical decision-tree models for parsing," pp. 276–283, 1995, doi: 10.3115/981658.981695.

[28] H. Fan, F. Xue, and H. Li, "Project-based as-needed information retrieval from unstructured AEC documents," *J. Manag. Eng.*, vol. 31, no. 1, pp. 1–11, 2014, doi: 10.1061/(ASCE)ME.1943-5479.0000341.

[29] K. Machová, F. Barčák, and P. Bednár, "A bagging method using decision trees in the role of base classifiers," *Acta Polytech. Hungarica*, vol. 3, no. 2, pp. 121–132, 2006.

BIOGRAPHIES OF AUTHORS

Layth Rafea Hazim is an assistant teacher at the Cisco Networking Academy, Tikrit University, Iraq. He received his BSc degree in Computer Science from Tikrit University in 2007, Msc degree from the Altinbas University, Turkey in 2018. He worked as a head of Elctronic Computer Cemter ECC at Tikrit University during the period 2020 until now.

Abdulrahman Ahmed Jasim is an assistant lectururer at Department of Network Engineering\Collage of Engineening, Al-Iraqia University. He received his BSc degree in Computer Engineering from Dijlah University in 2012, Msc degree in Elecrtrical and computer Engineering from Altinbas University, Turkey in 2018.

Wisam Dawood Abdullah received his B.Sc. degree in computer science from Tikrit University, Iraq and his M.S. degree in Information Technology (with concentration in Telecommunications and Networks) from the University Utara Malaysia (UUM). He received an expert certifications from Cisco Networking Academy CCNP, CCNA,CCNA Security, IoT, Entrepreneurship, Grid, Voice, Wireless Cloud, Linux, CCNA Cybersecurity and IT, also he is a NetAcad administrator in Cisco Networking Academy, Iraq, currently he is lecturer in the Tikrit University, Cisco Networking Academy, member in IEEE. Research interest: Protocol Engineering, Network Analysis, Internet Architecture and Technologies, Wireless Performances, Network Traffic Engineering, Data Mining, Future Internet, Internet of Things, AI, ML.