Materials Today: Proceedings xxx (xxxx) xxx

Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Classification of covid related articles using machine learning

Deepthi Godavarthi, Mary Sowjanya A

Dept. of CSSE, Andhra University College of Engineering (A), Visakhapatnam, AP, India

ARTICLE INFO

Article history:
Received 29 December 2020
Received in revised form 29 December 2020
Accepted 17 January 2021
Available online xxxx

Keywords:
COVID-19
Text classification
KNN
MLP
XGBoost
Explainability

ABSTRACT

Covid 19 pandemic has placed the entire world in a precarious condition. Earlier it was a serious issue in china whereas now it is being witnessed by citizens all over the world. Scientists are working hard to find treatment and vaccines for the coronavirus, also termed as covid. With the growing literature, it has become a major challenge for the medical community to find answers to questions related to covid-19. We have proposed a machine learning-based system that uses text classification applications of NLP to extract information from the scientific literature. Classification of large textual data makes the searching process easier thus useful for scientists. The main aim of our system is to classify the abstracts related to covid with their respective journals so that a researcher can refer to articles of his interest from the required journals instead of searching all the articles. In this paper, we describe our methodology needed to build such a system. Our system experiments on the COVID-19 open research dataset and the performance is evaluated using classifiers like KNN, MLP, etc. An explainer was also built using XGBoost to show the model predictions.

Selection and peer-review under responsibility of the scientific committee of the Emerging Trends in Materials Science, Technology and Engineering.

1. Introduction

The district named Wanzhou was affected by COVID 19 pandemic in China. Wanzhou became an enclosed center for epidemiological investigations with the lockdown of Wuhan and surrounding places on January 23, 2020. Hence an opportunity was provided for understanding the transmission dynamics and other risk factors associated with the spread of SARS-COV-2, the agent of COVID-19. 47 other Chinese cities also implemented the same measures to tackle COVID-19. Most of the COVID-19 cases are very mild in severity [1,2] thus reducing the likelihood that they would look around for testing and medical care [3].

CORD 19 corpora [4] introduced by the Allen Institute for AI and other research groups consists of over 200,000 scholarly articles, of which 100,000 are with full text, about COVID-19, SARS-CoV-2, and similar coronaviruses like SARS and MERS. People applied various AI-based techniques in information retrieval and NLP on this dataset for extracting important information. We propose a machine learning-based system that uses text classification application of NLP for extracting the information from this scientific literature. The main aim is to classify the covid related articles according to the journals in which they were published. Hence it can be considered as a multi-class classification problem.

2. Related works

On 7th January the raging virus was detected as coronavirus. It had >95% homology with the bat coronavirus and >70% similarity with the SARS-CoV. The covid related articles started increasing from February 2020 to date making it very difficult for human analysts to go through all the articles. Applications like identification of objects in images, speech to text conversion, news items matching, products related to user interests, etc can be done by using machine learning techniques [5]. They make use of the class of techniques named deep learning [6]. A system that is based on deep learning was proposed that uses NLP question answering methods to mine the literature [7]. CovidQA, a question-answering dataset was developed for covid 19 [8]. Shuja et al. [9] have formulated research domain taxonomy and identified features of datasets concerning the type, methods, application. Santos et al. [10] presented a dataset on COVID-19 in which research activities overview was provided so that it would be easy to find scientists and researchers who are active in the task of combating the disease. The models in data mining to predict COVID-19 patient's recovery using an epidemiological dataset of South Korea COVID-19 patients were implemented [11]. The assessment of information flow as well as, scientific collaboration quality was

https://doi.org/10.1016/j.matpr.2021.01.480
2214-7853/© 2021 Elsevier Ltd. All rights reserved.
Selection and peer-review under responsibility of the scientific committee of the Emerging Trends in Materials Science, Technology and Engineering.

performed which are important to find solutions for pandemic [12]. A dataset that consists of COVID-19 updates provided by the Nigeria Centre for Disease Control online from February to September [13] was provided. Saefi et al. [14] examined knowledge, practice, and attitude related to COVID-19 among undergraduate students in Indonesia and presented a dataset related to their examination. An automated theme-based visualization method that combines data modeling, information mapping, trend analysis was proposed [15]. Machine learning techniques are used for extracting activities and trends of covid related articles [16]. The author's gender distribution on covid related medical papers results is compared with published articles in the same journals in 2019 for articles from the US with first and last authors [17]. Kieuvongngam et al. [18] performed Text summarization on covid 19 using the Advances in pre-trained NLP models, BERT and OpenAI GPT-2. Chamola et al. [19] presented a deep review of important aspects related to covid 19 using the reliable source as well as used technologies such as IoT, Unmanned Aerial Vehicles (UAVs), blockchain, Artificial Intelligence (AI), and 5G to reduce covid 19 impacts. Feature extraction methods overview to recognize isolated or segmented characters were presented [20]. Dhole et al. [21] proposed a method where natural language interpretation and classification techniques are used for disease diagnosis. The usage of ROC curve for evaluating the performance of machine learning algorithms was investigated [22]. Machine learning techniques are used to illustrate the text classification process [23]. The overview of text classification algorithms was presented [24].

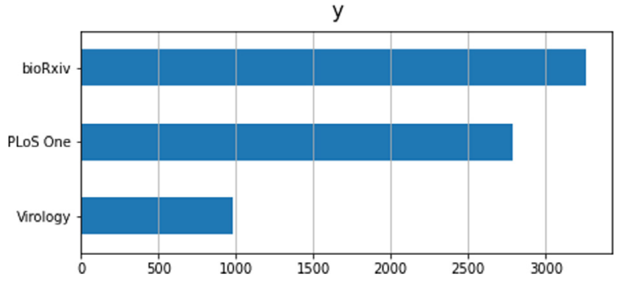


Fig. 2. Target.

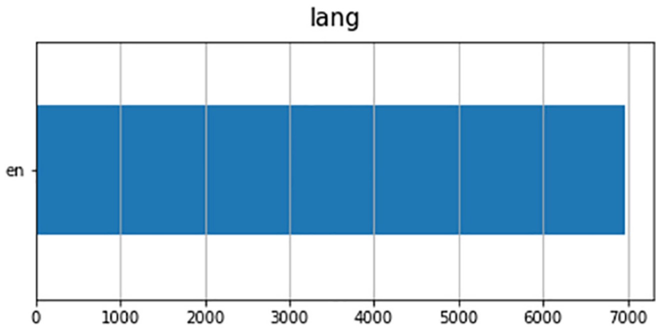


Fig. 3. Articles in English.

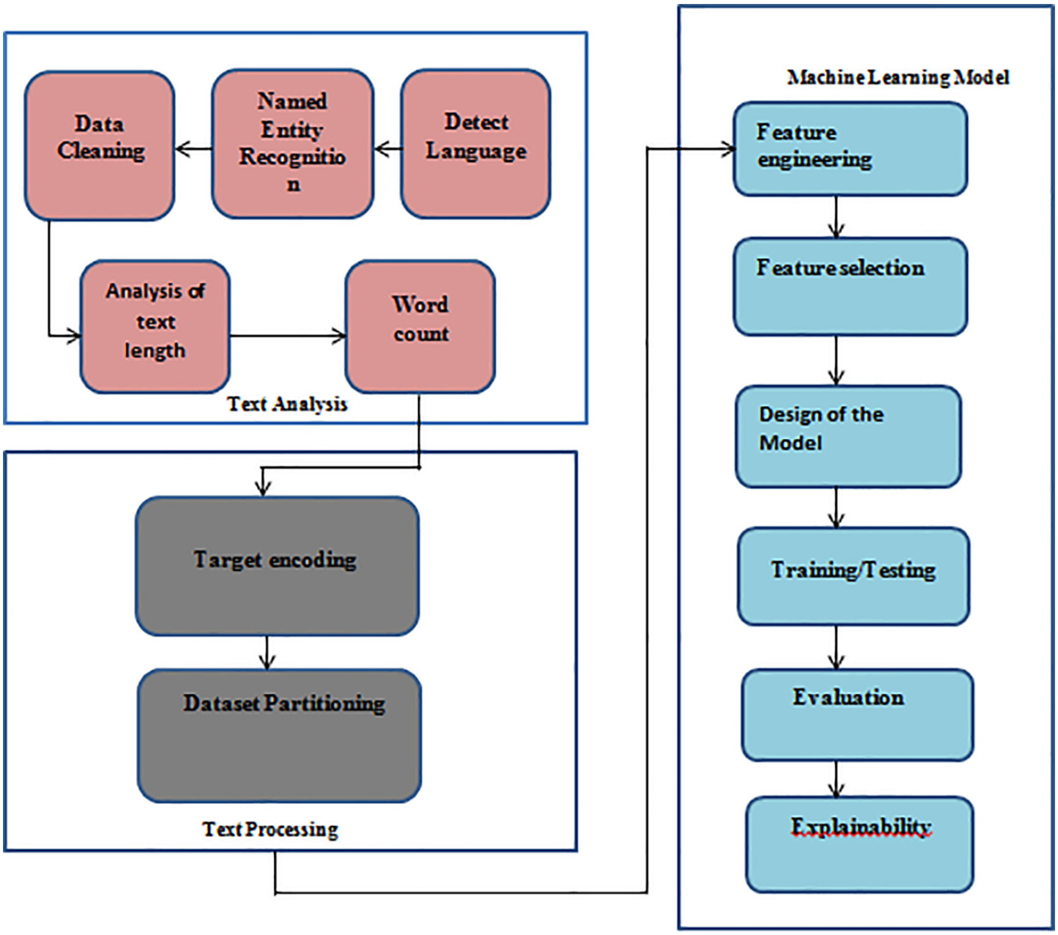
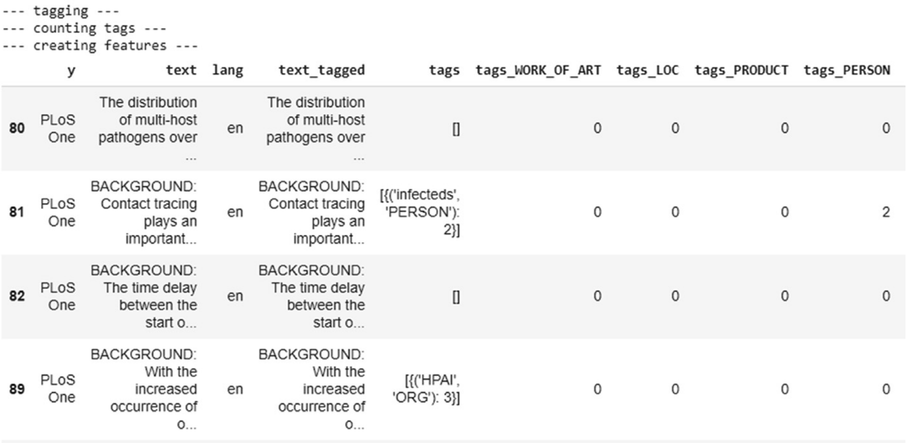


Fig. 1. System Architecture.



```
--- tagging ---
--- counting tags ---
--- creating features ---
```

	y	text	lang	text_tagged	tags	tags_WORK_OF_ART	tags_LOC	tags_PRODUCT	tags_PERSON
80	PLoS One	The distribution of multi-host pathogens over ...	en	The distribution of multi-host pathogens over ...	[]	0	0	0	0
81	PLoS One	BACKGROUND: Contact tracing plays an important...	en	BACKGROUND: Contact tracing plays an important...	[{('infecteds', 'PERSON'): 2}]	0	0	0	2
82	PLoS One	BACKGROUND: The time delay between the start o...	en	BACKGROUND: The time delay between the start o...	[]	0	0	0	0
89	PLoS One	BACKGROUND: With the increased occurrence of o...	en	BACKGROUND: With the increased occurrence of o...	[{('HPAI', 'ORG'): 3}]	0	0	0	0

Fig. 4. Named entity tagging with predefined categories.

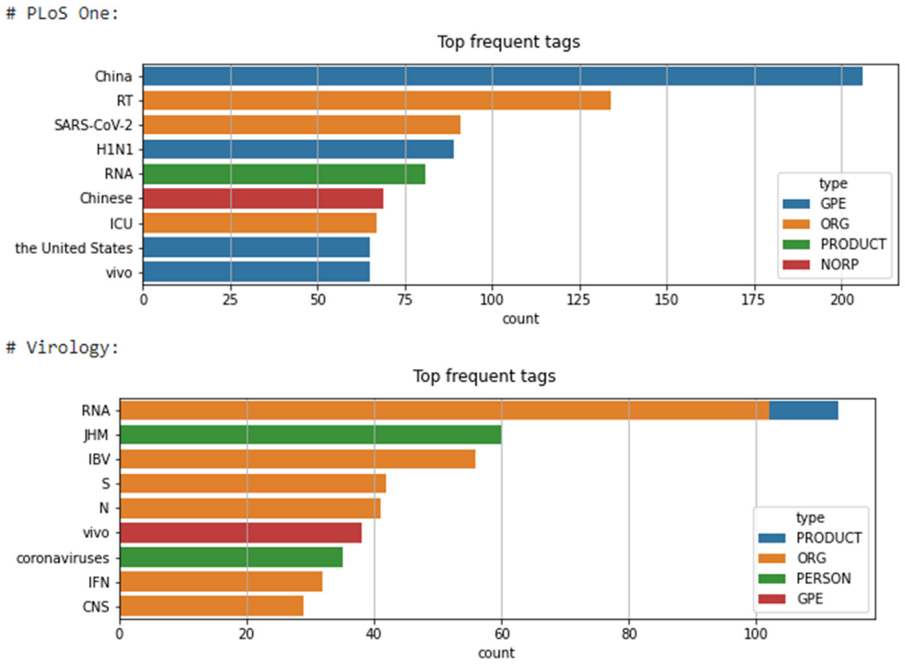


Fig. 5. (a) Top frequent tags in PLoS One journal: (b) Top frequent tags in Virology journal.

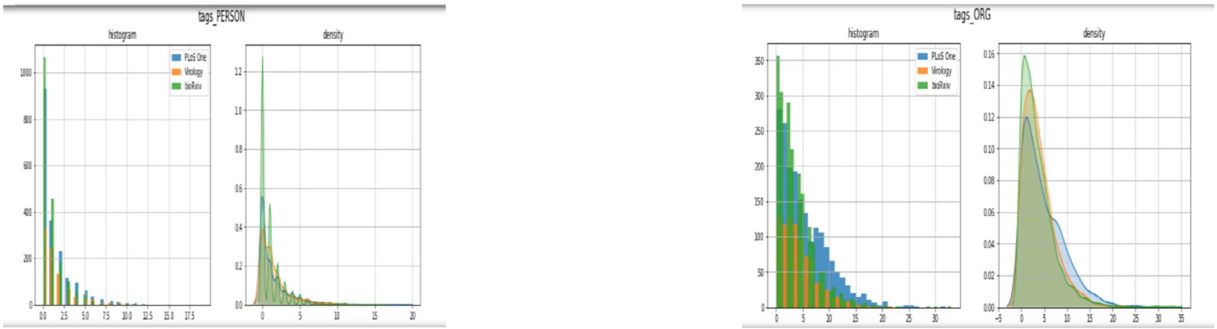
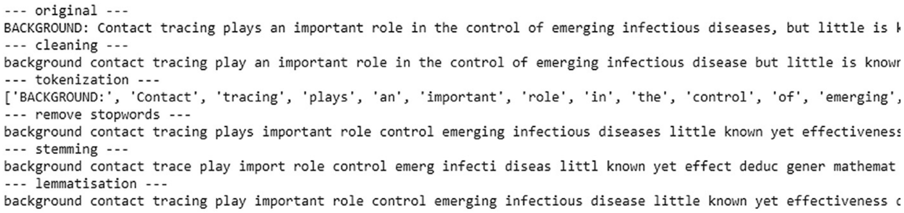


Fig. 6. (a) Histogram and density plots for entity 'PERSON': (b) histogram and density plots for entity 'ORG'.



```
--- original ---
BACKGROUND: Contact tracing plays an important role in the control of emerging infectious diseases, but little is k
--- cleaning ---
background contact tracing play an important role in the control of emerging infectious disease but little is know
--- tokenization ---
['BACKGROUND:', 'Contact', 'tracing', 'plays', 'an', 'important', 'role', 'in', 'the', 'control', 'of', 'emerging',
--- remove stopwords ---
background contact tracing plays important role control emerging infectious diseases little known yet effectiveness
--- stemming ---
background contact trace play import role control emerg infecti diseas littl known yet effect deduc gener mathemat
--- lemmatisation ---
background contact tracing play important role control emerging infectious disease little known yet effectiveness c
```

Fig. 7. Dataset Preprocessing.

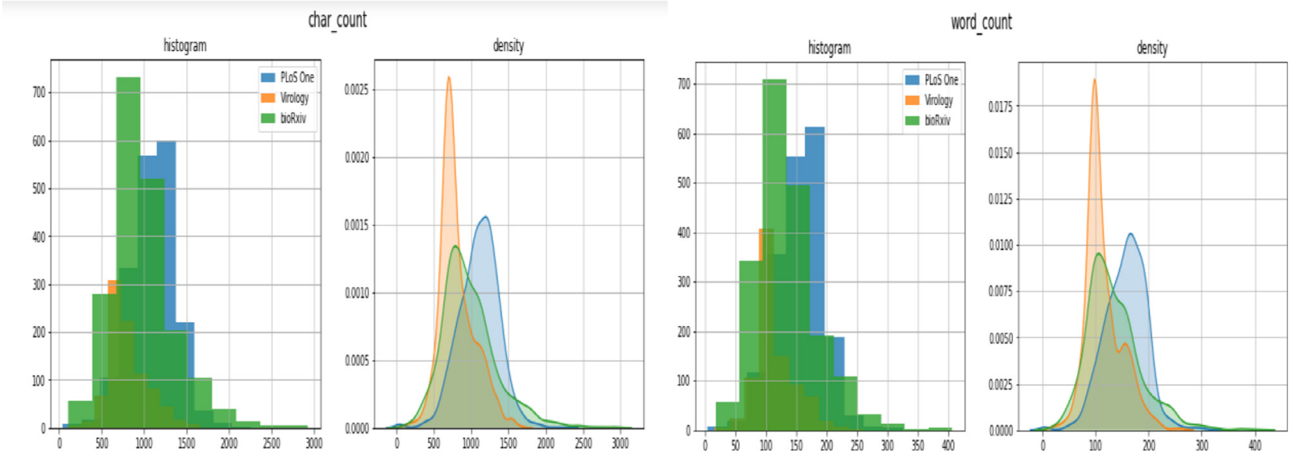


Fig. 8. (a) Plots obtained for char count: (b) plots obtained for word count.

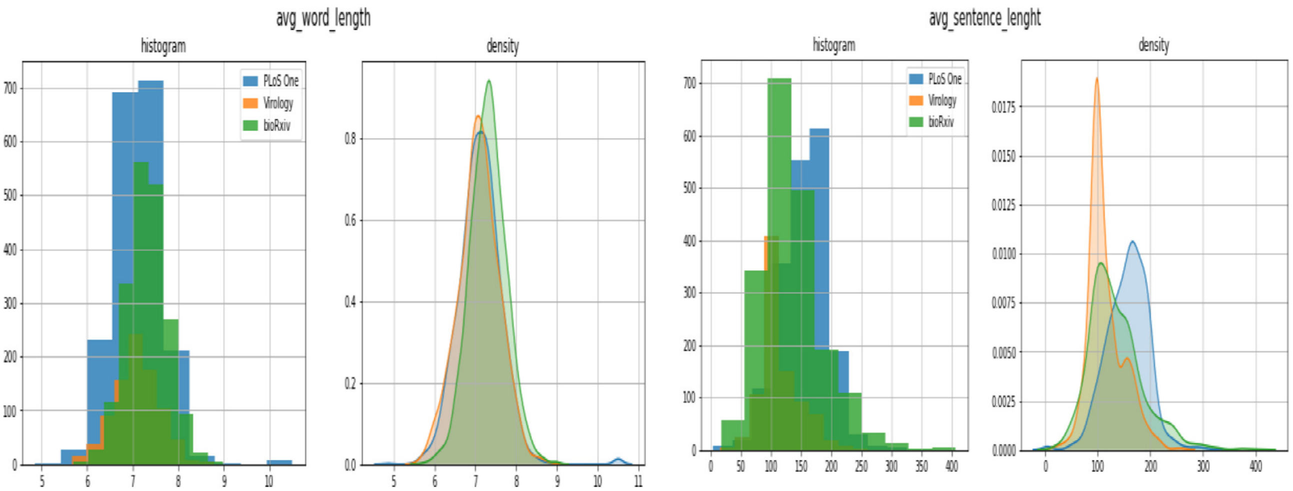
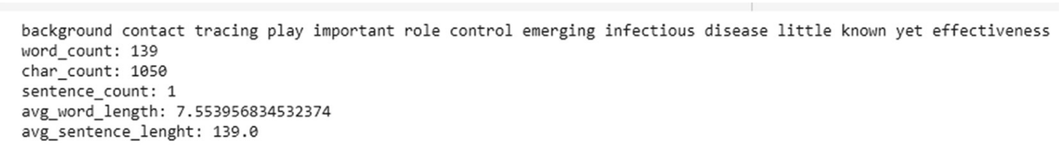


Fig. 9. (a) Plots obtained for avg word length: (b) plots obtained for avg sentence length.



```
background contact tracing play important role control emerging infectious disease little known yet effectiveness
word_count: 139
char_count: 1050
sentence_count: 1
avg_word_length: 7.553956834532374
avg_sentence_lenght: 139.0
```

Fig. 10. Text length analysis.

3. Methodology

Our system based on machine learning consists of 3 modules.1) Text analysis 2) Data Processing 3) Machine learning. The first module performs Language detection, Named entity recognition, Data cleaning, Length analysis, Word count. The second module performs Target encoding and Dataset partitioning. The third module performs Feature engineering with vectorizer, Feature selection, Model design, Train/test, Evaluate, Explainability. The architecture of our system is shown in Fig. 1.

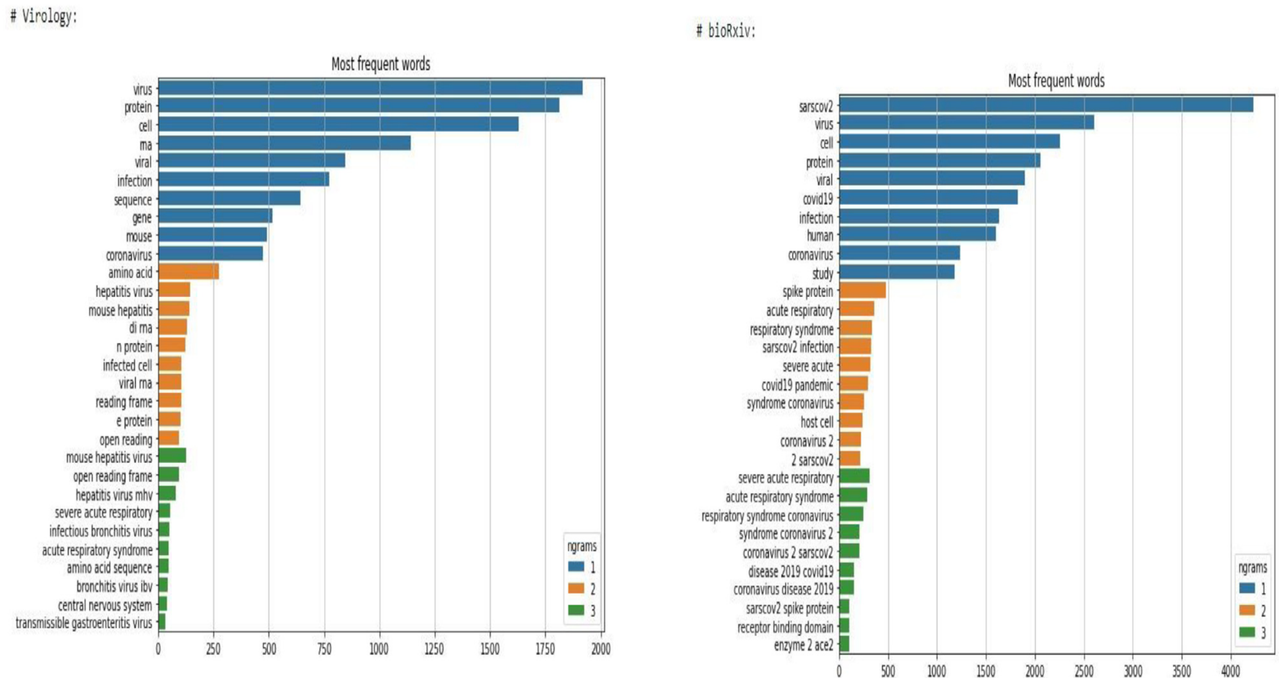


Fig. 11. (a) Most frequent words in virology journal: (b) Most frequent words in virology journal.

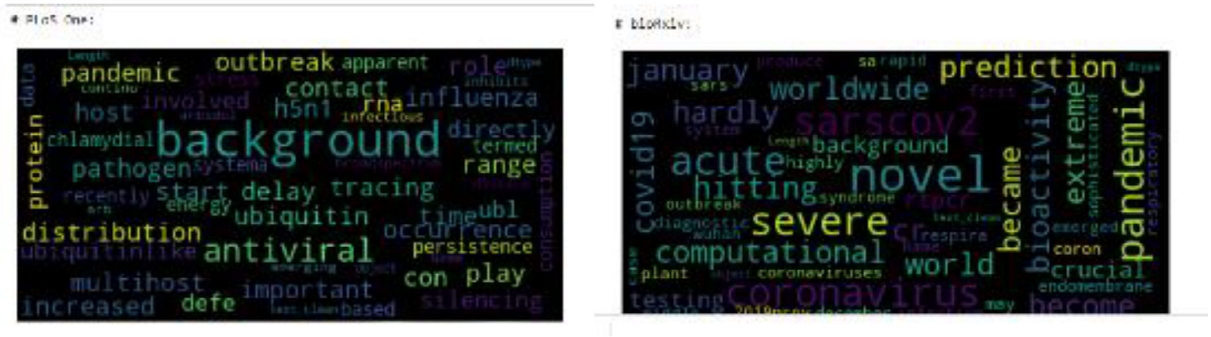
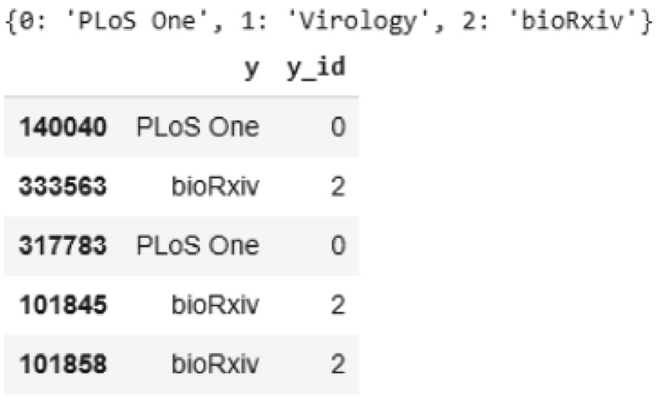


Fig. 12. (a) Word count for PLoS One journal: (b) Word count for bioRxiv journal.



```
{0: 'PLoS One', 1: 'Virology', 2: 'bioRxiv'}
```

	y	y_id
140040	PLoS One	0
333563	bioRxiv	2
317783	PLoS One	0
101845	bioRxiv	2
101858	bioRxiv	2

Fig. 13. Journals encoding.

3.1. Analysis of text

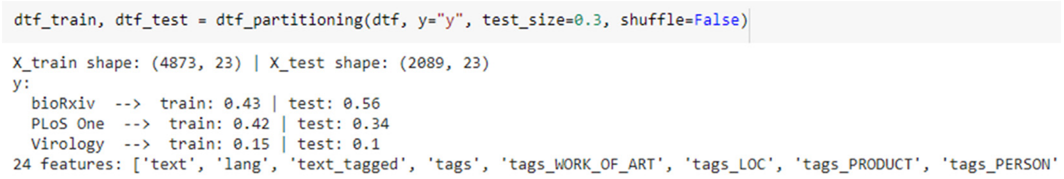
There are several journals in the dataset. In our model, we have considered a subset of 3 journals:bioRxiv, PLoS One, Virology represented in Fig. 2. The proportion of Virology is small compared to bioRxiv and PLoS One. To solve this issue we performed dataset resampling. First data cleaning was done followed by the extraction of insights from raw data. Then they were added as new columns in a data frame. This new information was used for the classification model.

3.1.1. Detect language

First, we are detecting the type of language in which the articles were published. Since there might be articles from multiple languages we are filtering out the articles in English. Langdetect package is used on articles of the dataset. It can be done on the complete dataset by adding a new column with information about the language as shown in Fig. 3.

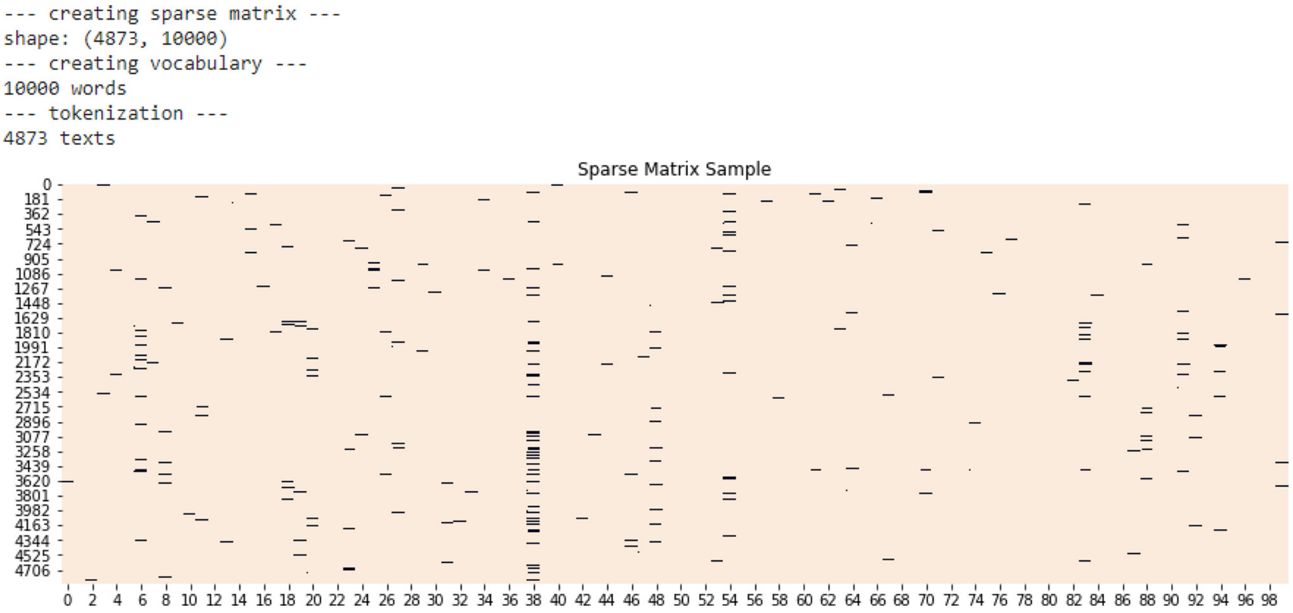
3.1.2. Named entity recognition

The process of tagging named entities present in raw text with predefined categories like names of persons, organizations, quantities, locations, etc is called NER. It takes a lot of time to train the NER model as it requires a rich dataset. So we use NER tools provided by SpaCy as it provides various NLP models to identify various entity categories. SpaCy model en_core_web_lg is used on abstracts. For each abstract, all the recognized entities are inserted



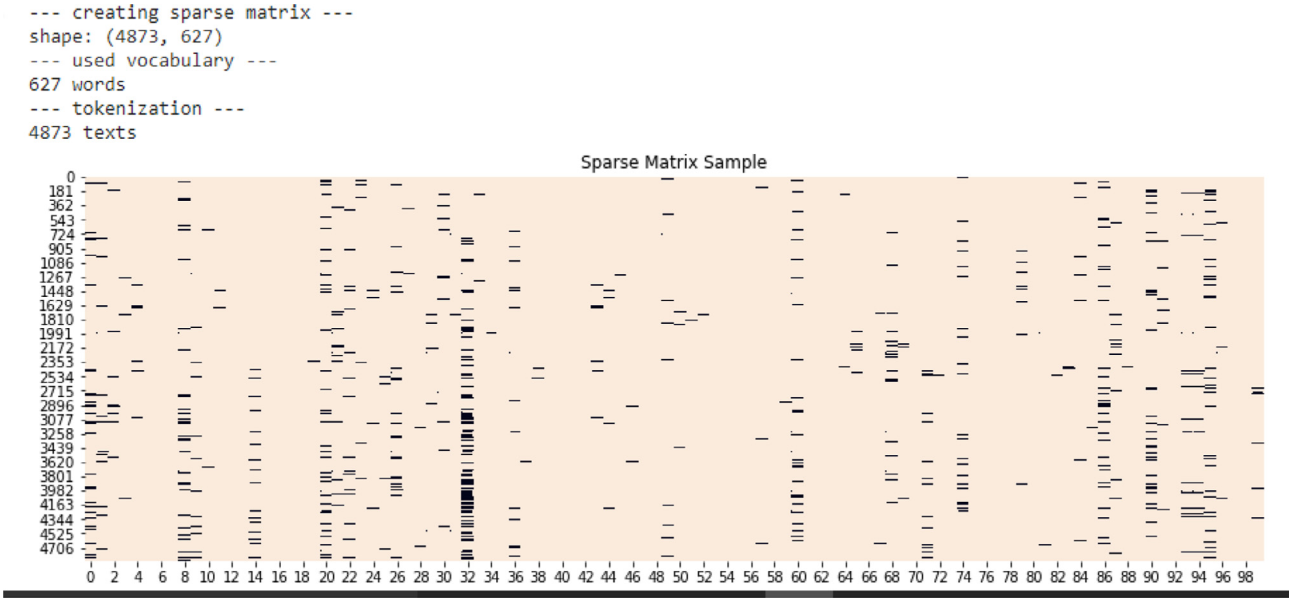
```
dtf_train, dtf_test = dtf_partitioning(dtf, y="y", test_size=0.3, shuffle=False)

X_train shape: (4873, 23) | X_test shape: (2089, 23)
y:
  bioRxiv  -->  train: 0.43 | test: 0.56
  PLoS One  -->  train: 0.42 | test: 0.34
  Virology  -->  train: 0.15 | test: 0.1
24 features: ['text', 'lang', 'text_tagged', 'tags', 'tags_WORK_OF_ART', 'tags_LOC', 'tags_PRODUCT', 'tags_PERSON'
```

Fig. 14. Dataset partitioning for model performance evaluation.



```
--- creating sparse matrix ---
shape: (4873, 10000)
--- creating vocabulary ---
10000 words
--- tokenization ---
4873 texts
```

Fig. 15. Feature matrix creation.

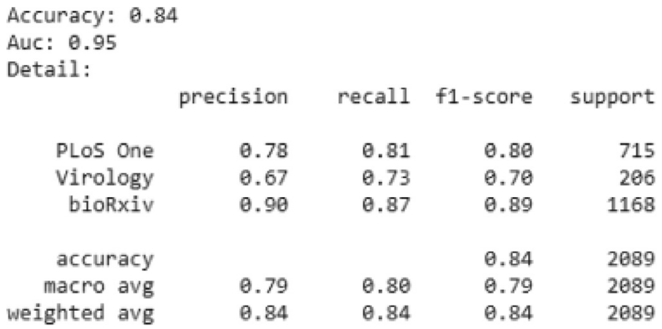


```
--- creating sparse matrix ---
shape: (4873, 627)
--- used vocabulary ---
627 words
--- tokenization ---
4873 texts
```

Fig. 16. Feature matrix obtained after dimensionality reduction.

into one column with a count giving the number of times that entity occurred in the text. Then one more column was created for every tag category and the count of each entity is place. We can observe the tag types distribution macro view as shown in Fig. 4. As Spacy can identify a person's name it was used for name detection and then the string would be modified.

The top frequent tags and histogram, density plots are shown in Figs. 5 and 6.



```
Accuracy: 0.84
Auc: 0.95
Detail:
              precision    recall  f1-score   support

    PLoS One       0.78      0.81      0.80       715
    Virology       0.67      0.73      0.70       206
     bioRxiv       0.90      0.87      0.89      1168

    accuracy                           0.84      2089
   macro avg       0.79      0.80      0.79      2089
weighted avg       0.84      0.84      0.84      2089
```

Fig. 17. The accuracy obtained on using XGBoost classifier.

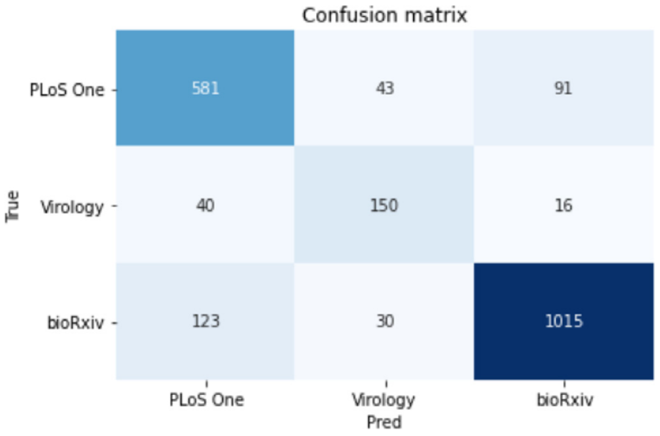


Fig. 18. Obtained confusion matrix for XGBoost.

3.1.3. Data cleaning

The raw data must be prepared in such a way that it must be suitable for the machine learning model to handle it. Steps in text cleaning depend on data type and the task required. Usually, strings will be converted to lower case letters and punctuations will be removed before splitting text into tokens. Since all tokens are not necessary we can remove tokens that don't give the required information. For example, words like "and", "the" are not useful as they occur in multiple places in the dataset. They are called stopwords and can be removed. While removing stopwords we must be very careful because if we remove the wrong token we may lose very important information. Word transformation techniques such as stemming and lemmatization are applied to produce words root form. All these preprocessing steps are written in one function and are applied to the complete dataset as shown in Fig. 7.

3.1.4. Analysis of length

We here identify whether one category is larger than the other since length could be considered as the only feature required for building a model. There are various ways to measure the length of text data.

Here the set of observations are divided into 3 samples based on the journal names (bioRxiv, PLoS One, virology), we compare the histograms and densities of the samples. The variable is said to be predictive if there are different distributions because there are different patterns for all groups. Though all 3 groups have a similar length distribution since they have different sizes density plots are essential and depicted as in Figs. 8 and 9. The text length analysis and the most frequent words in a particular journal are shown below in Figs. 10 and 11.

3.1.5. Word count

CountVectorizer from Scikit-learn is used to calculate word frequency. This vectorizer converts a set of documents into a matrix

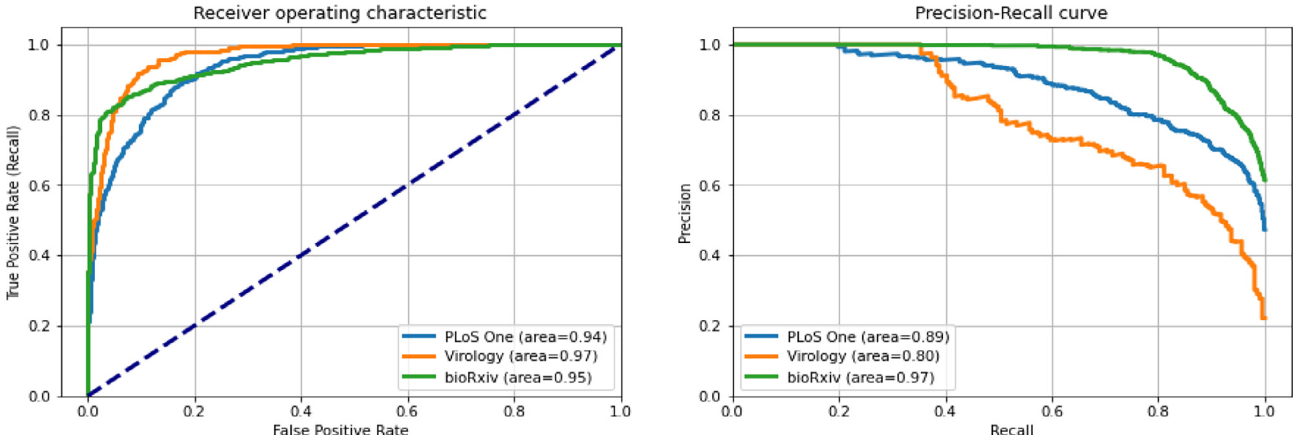
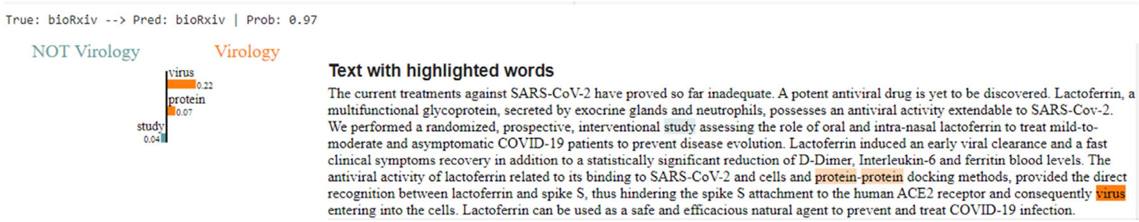


Fig. 19. (a) ROC curve for XGBoost: (b) Precision-Recall curve.



```
True: bioRxiv --> Pred: bioRxiv | Prob: 0.97
```

Text with highlighted words

The current treatments against SARS-CoV-2 have proved so far inadequate. A potent antiviral drug is yet to be discovered. Lactoferrin, a multifunctional glycoprotein, secreted by exocrine glands and neutrophils, possesses an antiviral activity extendable to SARS-Cov-2. We performed a randomized, prospective, interventional study assessing the role of oral and intra-nasal lactoferrin to treat mild-to-moderate and asymptomatic COVID-19 patients to prevent disease evolution. Lactoferrin induced an early viral clearance and a fast clinical symptoms recovery in addition to a statistically significant reduction of D-Dimer, Interleukin-6 and ferritin blood levels. The antiviral activity of lactoferrin related to its binding to SARS-CoV-2 and cells and protein-protein docking methods, provided the direct recognition between lactoferrin and spike S, thus hindering the spike S attachment to the human ACE2 receptor and consequently virus entering into the cells. Lactoferrin can be used as a safe and efficacious natural agent to prevent and treat COVID-19 infection.

Fig. 20. Explainer built using XGBoost to observe model predictions.

with a count of the token. This information can be visualized using the word cloud in which each tag frequency could be displayed with font size and color. The word count for various journals is shown below in Fig. 12.

3.2. Text processing

3.2.1. Target encoding

We have encoded variable ‘y’ that is assigned to journals and added a new column named ‘y_id’ to the data frame to perform journal encoding for clarity as shown in Fig. 13.

3.2.2. Dataset partitioning

Dataset is partitioned into a training set (70%) and test set (30%) to evaluate the performance of the model. So that the model can be fit on the training set as follows (Fig. 14).

3.3. Machine learning model

In Bag of words, vocabulary is built from document corpus and the number of times a word appears in the document is counted. A vector is used to represent a document with a length equal to the vocabulary size. Words in vocabulary act as features. If there are more documents, vocabulary size becomes larger which results in a huge feature matrix. To reduce this dimensionality problem, preprocessing was done. We can have common words with the highest frequency in the dataset but they may have little impact on the target variable so term frequency need not be considered as a good representation for text. Instead of normal counting, tf-idf can be used.

3.3.1. Feature engineering

Creating features from the raw text for a machine learning model is called feature engineering and is considered the most important phase of text classification. It is a process of information extraction from data for feature creation. Here we are using tf-idf vectorizer with 10,000 words limit along with capturing unigrams and bigrams. Now vectorizer is used on the preprocessed corpus of the train set for vocabulary extraction and feature matrix creation as shown in Fig. 15.

3.3.2. Feature selection

The feature matrix x_train has a shape of 4873(documents considered in training) * 10,000(vocabulary length). We can look for a word in vocabulary to know a certain word position. To reduce the dimensionality of a matrix we can drop a few unimportant columns by using feature selection where we will select a subset of only relevant variables.

As such the number of features has now been reduced from 10,000 to 627 by considering only similar features.

This new list of words can be given as input thus refitting the vectorizer on a corpus that produces a small feature matrix with less vocabulary. Fig. 16 shows the new feature matrix x_train with a shape of 4873*627.

3.3.3. Design of the model

Support vector machine, XGBoost and, MLP classifier are used to train the model, then predictions can be made based on knowledge about the related conditions. If the dataset is very large then this algorithm is very suitable because it takes each feature independently into consideration, calculating each category probability, and then the highest probability category will be predicted.

3.3.4. Training/testing

All the models are trained on the feature matrix and tested on the transformed test set and then a sci-kit learn pipeline was built. It is an application consisting of transformations and a final estimator list. Tfidf vectorizer and the model are kept in this pipeline so that it allows transformation and test data prediction.

Table 1
Performance comparison of our model with XGBoost, MLP, KNN.

Machine Learning Algorithms	Accuracy	AUC
XGBoost	0.84	0.95
MLP	0.83	0.94
KNN	0.76	0.86

3.3.5. Evaluation

The metrics such as Accuracy, Confusion Matrix, ROC, Precision, Recall, f1-score, support were used for evaluation as shown in Figs. 17 to 19. The performance of the bag of words model is evaluated. The BoW model got 76% of the test set correct (accuracy is 0.76) on using KNN and 83% on MLP classifier. XGBoost algorithm is used to improve the accuracy from 0.76 to 0.84.

3.3.6. Explainability

It is a process of explaining the internal mechanics of a system in human terms. As such an explainer lime package was used. The random observations from the test set were considered and model predictions can be observed. The words “virus”, “protein” pointed the model in the right direction (virology) as shown in Fig. 20.

It can be seen that XGBoost and MLP Classifier perform almost similarly on the dataset in Table 1.

4. Conclusions and future work

We have described our system consisting of 3 modules namely text analysis, processing and, machine learning model. Our system uses a CORD-19 dataset that consists of various scientific articles related to covid 19. We believe that by using our system it would be easy for the community to retrieve required articles from the journals of their interest and help them to combat the pandemic. In the future wish to develop a deep learning system with increased accuracy.

CRediT authorship contribution statement

Godavarthi Deepthi: Conceptualization, Methodology, Software, Visualization, Writing - original draft. **A. Mary Sowjanya:** Data curation, Supervision, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R.T. Gandhi, J.B. Lynch, C. del Rio, Mild or moderate Covid-19, N. Engl. J. Med. 383 (18) (2020) 1757–1766, https://doi.org/10.1056/nejmcp2009249.

[2] Z. Wu, J.M. McGoogan, Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72314 cases from the Chinese Center for Disease Control and Prevention, JAMA – J. Am. Med. Assoc. 323 (13) (2020) 1239–1242, https://doi.org/10.1001/jama.2020.2648.

[3] M. Peppa, W. John Edmunds, S. Funk, Disease severity determines health-seeking behaviour amongst individuals with influenza-like illness in an internet-based cohort, BMC Infect. Dis. 17 (1) (2017) 1–13, https://doi.org/10.1186/s12879-017-2337-5.

[4] L. L. Wang et al., CORD-19: The C, 2020.

[5] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, no. May, 2015, doi: 10.1038/nature14539.

[6] H. Wang, On the Origin of Deep Learning, pp. 1–72.

[7] D. Su, Y. Xu, T. Yu, F. Bin Siddique, E. J. Barezi, P. Fung, CAiRE-COVID: A Question Answering and Multi-Document Summarization System for COVID-19 Research, 2020, [Online]. Available: http://arxiv.org/abs/2005.03975.

[8] R. Tang et al., Rapidly Bootstrapping a Question Answering Dataset for COVID-19.

[9] J. Shuja, E. Alanazi, W. Alasmary, A. Alashaikh, COVID-19 open source data sets: a comprehensive survey, Appl. Intell. (2020), https://doi.org/10.1007/s10489-020-01862-6.

[10] B. S. Santos, I. Silva, M. da C. Ribeiro-Dantas, G. Alves, P. T. Endo, L. Lima, COVID-19: A scholarly production dataset report for research analysis, Data Br., vol. 32, 2020, doi: 10.1016/j.dib.2020.106178.

[11] L.J. Muhammad, M.M. Islam, S.S. Usman, S.I. Ayon, Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery, SN Comput. Sci., vol. 1, no. 4, 2020, doi: 10.1007/s42979-020-00216-w.

[12] J.H.I.K.D. Virag, Preliminary analysis of COVID - 19 academic information patterns: a call for open science in the times of closed borders, Scientometrics 124 (3) (2020) 2687–2701, https://doi.org/10.1007/s11192-020-03587-2.

[13] E. Ogundepo et al., An exploratory assessment of a multidimensional healthcare and economic data on COVID-19 in Nigeria, Data Br. 33 (2020), https://doi.org/10.1016/j.dib.2020.106424.

[14] M. Saefi et al., Survey data of COVID-19-related knowledge, attitude, and practices among indonesian undergraduate students, Data Br. 31 (2020), https://doi.org/10.1016/j.dib.2020.105855.

[15] P. Le Bras, A. Gharavi, D.A. Robb, A.F. Vidal, S. Padilla, M.J. Chantler, Visualising COVID-19 Research, no. May, pp. 1–11, 2020, [Online]. Available: http://arxiv.org/abs/2005.06380.

[16] S.K. Sonbhadra, S. Agarwal, P. Nagabhushan, Target specific mining of COVID-19 scholarly articles using one-class approach, pp. 1–12, 2020, [Online]. Available: http://arxiv.org/abs/2004.11706.

[17] J.P. Andersen, M.W. Nielsen, N.L. Simone, R.E. Lewiss, R. Jagsi, COVID-19 medical papers have fewer women first authors than expected, Elife 9 (734) (2020) 1–7, https://doi.org/10.7554/eLife.58807.

[18] V. Kieuvongngam, B. Tan, Y. Niu, Automatic Text Summarization of COVID-19 Medical Research Articles using BERT and GPT-2," 2020, [Online]. Available: http://arxiv.org/abs/2006.01997.

[19] V. Chamola, V. Hassija, V. Gupta, M. Guizani, A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT , Drones, AI , Blockchain, and 5G in Managing Its Impact, no. April, pp. 90225–90265, 2020.

[20] Feature extraction methods for character recognition – a survey, vol. 29, no. 4, pp. 641–662, 1996.

[21] G. Dhole, N. Uke, NLP Based Retrieval of Medical Information for Diagnosis of Human Diseases, pp. 243–248, 2014.

[22] P. Recognition, A.E. Bradley, The use of the Area Under The ROC Curve in the Evaluation Of machine learning Algorithms, vol. 30, no. 7, pp. 1145–1159, 1997.

[23] M. Ikonomakis, S. Kotsiantis, V. Tampakas, Text classification using machine learning techniques, WSEAS Trans. Comput. 4 (8) (2005) 966–974, https://doi.org/10.11499/sicejl1962.38.456.

[24] K. Kowsari, K.J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, D. Brown, Text classification algorithms: a survey, Inf. 10 (4) (2019) 1–68, https://doi.org/10.3390/info10040150.