Indian Journal of Biochemistry & Biophysics
Vol. 57, October 2020, pp. 602-605

Machine learning approach for COVID-19 crisis using the clinical data

NRP Kumar* & NS Shetty

Jain Deemed to be University, Bangalore-560 069, Karnataka, India

Received 18 September 2020; *revised 23 September 2020*

We try to identify the impact of innovation headways and its rapid affect in each field of life, be it clinical or some other field; computerized reasoning deployed the prominent approach for indicating the authenticated outcomes in the field of medical services through its dynamic nature in investigating the information. COVID-19 has influenced all the nations around the globe in a short period of time duration; Individuals everywhere over the world are defenceless, against its results in the future. It is necessary to build up a control framework that will distinguish the Covid. One of the answers for control the flow ruin can be the conclusion of illness with the assistance of different artificial intelligence instruments.

In this paper, we ordered literary clinical reports into four classes by utilizing old style and troupe AI calculations. Feature designing was performed utilizing procedures like Term recurrence/reverse archive recurrence (TF/IDF), Bag of words (BOW) and report length. These highlights were provided to customary and troupe AI classifiers. Calculated relapse and Multinomial Naive Bayes demonstrated preferred outcomes over other ML calculations by having 96.2% testing exactness. In the future intermittent neural organization can be utilized for better exactness.

Keywords: Accentuation lemmatisation, Bagging, Dyspnoea

In October 2019, the novel coronavirus was identified in the Wuhan city of China[1] and was accounted to the World Health Organization (W.H.O) on 31st December 2019. The infection created a worldwide disaster and was named as COVID-19 by W.H.O on eleventh February 2020[1]. The COVID-19 is the group of infections including SARS, ARDS. W.H.O proclaimed this episode as a general wellbeing crisis[2] and referenced the accompanying; the infection is being sent through the respiratory plot when a solid individual comes in con-affability with the tainted individual. The infection may communicate between people through different roots which are presently muddled. The tainted individual shows manifestations inside 2–14 days, contingent upon the brooding time of the centre east respiratory disorder (MERS), and the extreme intense respiratory condition (SARS). As per WHO the signs and manifestations of mellow to direct cases are dry hack, exhaustion, and fever while as in extreme cases dyspnoea (windedness), Fever and sleepiness may happen[3,4].

Further, Figure 1 illustrates the covid density map as of August 2020. Also, the people having different infections like asthma, diabetes, and coronary illness are more defenceless against the infection and may turn out to be seriously sick. The individual is analysing dependent on indications and his movement history. Fundamental signs are being watched acutely of the customer having manifestations. No particular treatment has been found on 10th April 2020, and patients are being dealt with apparently. The information gave by John Hopkins University as X-beam pictures and different specialists fabricate a model of AI that groups X-beam pictures into COVID-19. Since the most recent information distributed by Johns Hopkins gives the metadata of these pictures.

The information comprises of clinical reports as text in this paper, we are characterizing that text into four unique classes of illnesses with the end goal that it can help in recognizing COVID-19 from prior clinical indications. We utilized regulated AI methods for characterizing the content into four different classes COVID, SARS, ARDS, and Both (COVID, ARDS). We are additionally utilizing group learning procedures for order. Area 2 gives the writing review with respect to the proposed work. The system for distinguishing Covid from clinical content information is being discussed in Sects. 3 and 4 give the exploratory consequences of the proposed system and Sect. 5 finishes up our work.

Related Work

AI and regular natural language processing utilize large information- based models for design acknowledgment, clarification, and forecast. NLP has

*Correspondence:
E-mail: swarnaputra@gmail.com

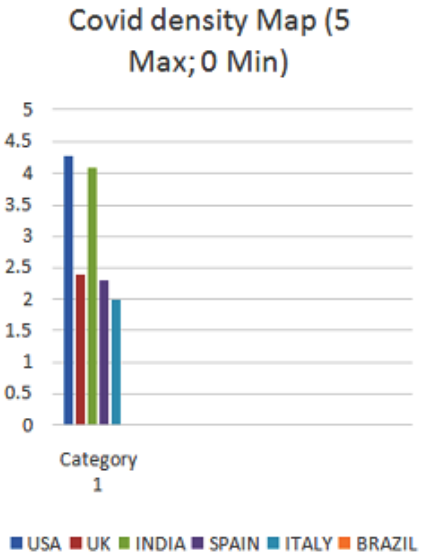

Fig. 1 — Covid density map

increased a lot of enthusiasm for late years, generally in the field of text investigation; Classification is one of the significant errands in text mining and can be performed utilizing various Liu W[6]. Khanday AMUD[7] played out a SWOT examination of different directed and solo content grouping calculations for mining the unstructured information. The different uses of text orders are supposition investigation, extortion recognition, and spam identification *etc*. Conclusion mining is significantly being utilized for races, promotion, and business, and so on. Kumar[8] investigated the sentiments of Indian government ventures with the assistance of the vocabulary-based word reference.

These reasons can be valuable to analyse and anticipate COVID-19. Firm and careful conclusion of COVID-19 can spare a large number of lives and can create a gigantic measure of information on which an AI (ML) models can be prepared. ML may give a valuable contributions to this respect, specifically in making analyse dependent on clinical content, radiography Images, and so forth. As per Bullock[11], Machine learning and profound learning can supplant people by giving a precise finding. The ideal analysis can spare radiologists' time and can be cost-effective than standard tests for COVID-19. X-beams and figured tomography (CT) sweeps can be utilized for preparing the AI model. A few activities are in progress in such a manner. Wang and Wong[12] created COVID-Net, which is a profound convolutional neural organization, which can analyze COVID-19 from chest radiography pictures. When the COVID-19 is identified in an individual, the inquiry is whether and how seriously that individual will be influenced. Not all COVID-19 positive patients will require thorough consideration. Having the option to anticipation who will be influenced all the more seriously can help in coordinating and arranging clinical asset assignment and use. Yan[13] utilized AI to build up a prognostic forecast calculation to foresee the mortality danger of an individual that has been tainted, utilizing information from (just) 29 patients at Tongji Hospital in Wuhan, China. Jiang[14] proposed an AI model that can anticipate an individual influenced by COVID-19 and has the likelihood to create intense respiratory trouble disorder (ARDS). The proposed model brought about 80% of exactness. The examples of 53 patients were utilized for preparing their model and are confined to two Chinese emergency clinics. ML can be utilized to analyze COVID-19 which needs a great deal of examination exertion however isn't yet broadly operational. Since less work is being done on conclusion and foreseeing utilizing text, we utilized AI and gathering learning models to order the clinical reports into four classifications of infections.

Methodology

The proposed technique comprises of 2.1 to 2.5 advances. In sync 2.1 information assortment is being performed and 2.2 characterizes conventional AI calculations are examined, and 2.5 gives a review of gathering AI calculations. The visual portrayal of the proposed strategy is being examined beneath.

Data collection

W.H.O announced the Coronavirus pandemic as Health Emergency. The analysts and medical clinics give open admittance to the information with respect to this pandemic. We gathered from an open-source information storehouse GitHub.1 In which around 200 patient's information is put away which have demonstrated side effects of Covid and different infections. The information comprises of around 24 ascribes specifically understanding id, balanced, sex, age, discovering, endurance, intubated, went_icu, needed_supplemental_O2, extubated, temperature, pO2_saturation, leukocyte_count, neutrophil check, lymphocyte tally, see, methodology, date, area, envelope, filename, DOI, URL. Permit. Clinical notes and different notes.

Relevant dataset

Our work is with respect to message mining so we extricated clinical notes and discoveries. Clinical notes comprise of text while the quality discovering

comprises the name of the corresponding text. Around 200 reports were utilized and their length was determined. We consider just those reports that are written in the English language. Further the length dispersion of clinical reports that are written in English. The clinical reports are named to their corresponding classes. In our dataset, we have four classes COVID, ARDS, SARS, and Both (COVID, ARDS). The various classes where in clinical content is being sorted and comparing report length.

Pre-processing

The content is unstructured so it should be refined with the end goal that AI should be possible. Different advances are being followed in this stage; the content is being cleaned by eliminating superfluous content. Accentuation and lemmatisation are being done with the end goal that the information is refined in a superior manner. Stop words, images, URL's, joins are taken out with the end goal that characterization can be accomplished with better exactness.

Feature building

From the pre-processed clinical reports, different highlights are removed according to the semantics and are changed over into the refining of information, 2.3 gives a diagram of pre-process.

Results and Discussion

We utilized the windows framework with 4 GB Ram and 2.3 GHz processors for playing out this work. Sciket learn apparatus is being utilized for performing AI characterization with the assistance of different libraries like NLTK, STOP-WORDS, and so forth for improving the exactness of all the AI calculations pipeline is being utilized. After performing the factual calculation, more profound bits of knowledge about the information was accomplished. The information is being parted into 70:30 proportion where 70% information is being utilized for preparing the model and 30% is utilized for testing the model. We have clinical content reports of 2 patients that are marked into four classes. The grouping was finished utilizing AI algorithms by providing them includes that were separated in the element building step. To investigate the generalization of our model from preparing information to inconspicuous information and lessen the chance of overfitting, we split our underlying dataset into independent preparing and test subsets. The ten times cross-approval system was directed for all calculations, and this cycle was rehashed multiple times freely to keep away from the examining predisposition presented by haphazardly partitioning the dataset in the cross-approval. Table 1 gives a near examination of all the old -style AI strategies that are utilized for playing out this undertaking. Table 1 gives a near investigation of all the old-style AI and Ensemble learning techniques that are utilized for playing out the errand of grouping the clinical content into four classes. The outcomes demonstrated that strategic relapse and Multinomial Naıve Bayes Algorithm shows preferred outcome overall different calculations by having an accuracy 94.3%, recall 95%, F1 score 92%,and precision 90% different calculations like irregular woodland, gradient boosting likewise demonstrated great outcomes by having exactness 91.3% individually. The visualized relative investigation of the apparent multitude of calculations that are utilized in our work has appeared in (Fig. 1). Since we as a whole know, the COVID-19 information is least accessible. To get the genuine exactness of the model we tested it in two phases. In the principal stage, we took 75% of the accessible information and it shows less precision when contrasted with the phase wherein entire information was utilized for experimentation. So we can reason that if more information is provided to these calculations, there are odds of progress in performance. As we are confronting a serious test in handling the dangerous infection, our work will in one way or another assistance the network by breaking down the clinical reports and take important activities. Likewise, it was dissected that the COVID-19 patients report length is a lot littler than different classes and it ranges from 125 characters to 350 characters.

Table 1 — Machine learning algo comparison

Algorithm	Precision	Recall	F1 score	Accuracy (%)
Logistic regression	0.90	0.95	0.92	94.3
Multinomial Naıve Bayesian	0.90	0.95	0.92	94.3
Support vector machine	0.82	0.91	0.86	90.6
Decision tree	0.90	0.90	0.90	90.8
Bagging	0.92	0.92	0.92	92.5
Adaboost	0.85	0.91	0.88	90.6
Random forest	0.89	0.92	0.91	91.3
Stochastic gradient boosting	0.89	0.92	0.91	91.3

Conclusion

In the present study of novel coronavirus researchers trying to come with an effective vaccine, during the process of deploying the vaccine it undergoes many trail phases and it might take huge time for a better outcome. To overcome from the present situation, incorporating

computational features like machine learning will safeguard the human lives.

We used around 200 clinical trials namely COVID, SARS, ARDS, and Both. Various features are being used as a bag of words. Machine learning algorithms like logistic regression, multinomial naïve Bayesian classifier. After performing the analysis and we could conclude that logistic regression and multinomial naïve Bayesian classifier had given a good result by having 90% precision, 95% recall, 92% f1 score, and accuracy 94.3%. Even the other algorithm that should good result is the random forest, stochastic gradient boosting, decision trees, and boosting. The efficiency of the accuracy, precision, recall, f1 score can be improved by providing a larger data. The virus can be also classified based on gender, to know whether Male is more infected than female we need information for that. We can apply more engineering and deep learning technique for better understanding.

Conflict of Interest

All authors declare no conflict of interest.

References

1 Pathak M, COVID-19 research in India: A quantitative analysis. *Indian J Biophys Biophys*, 57 (2020) 352.

2 World Health Organisation, Coronavirus disease (COVID19) situation report, (https://www.who.int/docs/defaultsource/coronaviruse/situation-reports/20200512-COVID-19- sitrep-113.pdf? sfvrsn=feac3b6d_2) Accessed on 12th May 2020.

3 Dimensions COVID-19 publications, datasets and clinical trials. Dimensions. Dataset. https://doi.org/10.6084/m9.figshare.11961063.v21 (accessed on 14th May 2020).

4 Lou J, Tian SJ, Niu SM, Kang XQ, Lian HX, Zhang LX & Zhang JJ, Coronavirus disease 2019: A bibliometric analysis and review. *Eur Rev Med Pharmacol Sci*, 24 (2020) 3411.

5 Kousha K & Thelwall M, COVID-19 publications: Database coverage, citations, readers, tweets, news, Facebook walls, Reddit posts, Quantitative science studies, (2020) (in press).

6 Liu W, Morse JS, Lalonde T & Xu S, Learning from the past: possible urgent prevention and treatment options for severe acute respiratory infections caused by 2019-nCoV. *Chem Biochem*, 21 (2020) 730.

7 Khanday AMUD, Amin A, Manzoor I & Bashir R, ‘‘Face Recognition Techniques: A Critical Review’’2018.

8 Kumar A, Dabas V & Hooda P, Text classification algorithms for mining unstructured data: a SWOT analysis. *Int J Inf Technol*, (2018) https://doi.org/10.1007/s41870-017-0072-1.

9 Verma P, Khanday AMUD, Rabani ST, Mir MH & Jamwal S Twitter Sentiment Analysis on Indian Government Project using R. *Int J Recent Tech Eng*,8 (2019) 8338.

10 Chakraborti S, Choudhary A & Singh A, A machine learning based method to detect epilepsy. *Int J Inf Technol*, 10 (2018) 257.

11 Bullock J, Luccioni A, Pham KH, Lam CSN & Luengo-Oroz M, Mapping the landscape of artificial intelligence applications against COVID-19. (2020) https://arxiv.org/abs/2003.11336v1.

12 Wang L & Wong A, COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 Cases from chest radiography images. (2020) https://arxiv.org/abs/2003.09871.

13 Yan L, Zhang HT, Xiao Y, Wang M, Sun C, Liang J, Li S, Zhang M, Guo Y, Xiao Y, Tang X, Cao H, Tan X, Huang N, AMD A, Luo BJ, Cao Z, Xu H & Yuan Y, Prediction of criticality in patients with severe COVID-19 Infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. *medRxiv*, (2020) https://doi.org/10.1101/2020.02.27. 20028027.

14 Jiang X, Coffee M, Bari A, Wang J, Jiang X, Huang J, Shi J, Dai J, Cai J, Zhang T, Wu Z, He G & Huang Y, Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *CMC-Comput Mater Con*, 63 (2020) 537.

15 Description of Logistic Regression Algorithm. https://machine learningmastery.com/logistic-regression-for-machine-learning/. Accessed 15 May 2019.

16 Description of Multinomial Naive Bayes Algorithm https://www.3pillarglobal.com/insights/document-classification-using-multinomial-naive-bayes-classifier. Accessed 15 May 2019.

17 Khanday AMUD, Khan QR & Rabani ST, SVMBPI: support vector machine based propaganda identification. *SN Appl Sci*, (accepted).

18 Description of Decision Tree Algorithm: https://dataspirant.com/2017/01/30/how_decision_tree_algorithm_works/. Accessed 10 July 2019.

19 Description of Boosting Algorithm: https://towardsdatascience.com/boosting. Accessed 10 July 2019.

20 Description of Adaboost Algorithm: https://towards datascience.com/boosting-algorithm-adaboost-b673719ee60c. Accessed 10 July2019.

21 Katuwal R & Suganthan PN, Enhancing Multi-Class Classification of Random Forest using Random Vector Functional Neural Network and Oblique Decision Surfaces, (2018) Arxiv:1802.01240v1.

22 Friedman JH, Stochastic gradient boosting. *Comput Stat Data Anal*, 38 (2002) 367.