JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PERFORMANCE OF SEVERAL MACHINE LEARNING ALGORITHMS FOR DETECTING COVID-19 ON CLINICAL TEXT DATA

ESWARDU GANDIKOTA [#1] , B.NANDAN KUMAR [#2],D.D.D SURI BABU [#3]

[#1] M.Tech Student, Department of Computer Science and Engineering,
DNR College of Engineering and Technology, Sri RamaPuram, Balusumudi,
Bhimavaram - 534202.
[#2] Assistant Professor, Department of Computer Science and Engineering,
DNR College of Engineering and Technology, Sri RamaPuram, Balusumudi,
Bhimavaram - 534202.
[#3] HOD & Associate Professor, Department of Computer Science and Engineering,
DNR College of Engineering and Technology, Sri RamaPuram, Balusumudi,
Bhimavaram - 534202.

ABSTRACT

Innovation progressions rapidly affect each field of life, be it clinical field or some other field. Man-made consciousness has shown the promising outcomes in medical services through its dynamic by examining the information. Coronavirus has influenced in excess of 100 nations in an issue of no time. Individuals everywhere on the world are helpless against its results in future. It is basic to foster a control framework that will identify the Covid. One of the answer for control the flow destruction can be the conclusion of infection with the assistance of different AI instruments. In this paper, we characterized literary clinical reports into four classes by utilizing traditional and troupe AI calculations. Highlight designing was performed utilizing methods like Term recurrence/backwards archive recurrence (TF/IDF), Bag of words (BOW) and report length. These highlights were provided to customary and outfit AI classifiers. Calculated relapse and Multinomial Naïve Bayes showed preferred outcomes over other ML calculations by having 96.2% testing exactness. In future intermittent neural organization can be utilized for better exactness.

Keywords: Artificial Intelligence, COVID-19, Machine Learning, Ensemble Model.

1. INTRODUCTION

In December 2019, the novel Covid showed up in the Wuhan city of China [1] and was accounted for to the World Health Organization (W.H.O) on 31st December 2019. The infection made a worldwide danger 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 announced this flare-up as a general wellbeing crisis [2] and referenced the accompanying; the infection is being sent through the respiratory lot when a sound individual interacts with the tainted individual. The infection may communicate between people through different roots which are presently indistinct. The contaminated individual shows indications inside 2–14 days, contingent upon the brooding time of the center east respiratory condition (MERS), and the extreme intense respiratory disorder (SARS).

As per W.H.O the signs and indications of gentle to direct cases are dry hack, weakness and fever while as in extreme cases dyspnea (windedness), Fever and sluggishness may happen [3, 4]. The people having different illnesses like asthma, diabetes, and coronary illness are more powerless against the infection and may turn out to be seriously sick. The individual is analyze dependent on indications and his movement history. Indispensable signs are being noticed definitely of the customer having side effects. No particular treatment has been found as on tenth April 2020, and patients are being dealt with apparently. The medications like hydroxychloriquine, antipyretic, against virals are utilized for the indicative treatment. As of now, no such antibody is created for forestalling this lethal illness, and we may avoid potential risk to forestall this sickness. By washing hands consistently with cleanser for 20 s and staying away from close contact with others by keeping the distance of around 1 m may decrease the odds of getting influenced by this infection. While sniffling, Covering the mouth and nose with the assistance of dispensable tissue and staying away from the contact with the nose, ear and mouth can help in its counteraction.

SARS is an airborne infection that showed up in 2003 in China and influenced 26 nations by having 8 K cases around the same time and moved from one individual to another. The signs and side effects of SARS are fever, cold, looseness of the bowels, shuddering, disquietude, myalgia and dyspnea. The ARDS (intense respiratory trouble disorder) is described by fast beginning of irritation in lungs which prompts respiratory disappointment and its signs and manifestations are pale blue skin tone, weakness and windedness. ARDS is analyzed by PaO2/FiO2 proportion of under 300 mm Hg. Till tenth of April 2020; practically 1.6 million affirmed instances of Covid are distinguished all throughout the planet. Just about 97 K people have kicked the bucket and 364 K people have recuperated from this lethal infection [5]. Figure 1 shows the overall information in regards to Covid. Since no medication or immunization is made for restoring the COVID-19. Different paramedical organizations have asserted of fostering an antibody for this infection. Less testing has additionally brought about this infection as we do not have the clinical assets because of pandemic. Since a great many are being tried positive step by step all throughout the planet, it is unimaginable to expect to test every one of the people who show indications

Apart from clinical techniques, AI gives a great deal of help in recognizing the sickness with the assistance of picture and text based information. AI can be utilized for the ID of novel Covid. It can likewise gauge the idea of the infection across the globe. Notwithstanding, AI requires an immense measure of information for ordering or foreseeing sicknesses. Directed AI calculations need commented on information for ordering the content or picture into various classes. From the previous decade, an enormous measure of progress is being made in this space for settling some basic activities. Ongoing pandemic has drawn in numerous specialists all throughout the planet to tackle this issue. Information given by John Hopkins University as X-beam pictures and different scientists fabricate a model of AI that groups X-beam picture into COVID-19 or not. 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 grouping that text into four unique classifications of illnesses with the end goal that it can help in recognizing Covid from prior clinical manifestations. We utilized administered AI strategies for ordering the content into four unique classifications COVID, SARS, ARDS and Both (COVID, ARDS). We are additionally utilizing group learning methods for order.

2. LITERATURE SURVEY

Literature survey is that the most vital step in the software development process. Before developing the new application or model, it's necessary to work out the time factor, economy, and company strength. Once all these factors are confirmed and got approval then we can start building the application. The literature survey is one that mainly deals with all the previous work which is done by several users and what are the advantages and limitations of those previous models. This literature survey is mainly used for identifying the list of resources to construct this proposed application.

MOTIVATION

AI and common language preparing utilize enormous information based models for design acknowledgment, clarification, and forecast. NLP has acquired a lot of interest lately, generally in the field of text examination, Classification is one of the significant assignment in text mining and can be performed utilizing various calculations [6].

Kumar et al. [7] played out a SWOT investigation of different managed and unaided content arrangement calculations for mining the unstructured information. The different utilizations of text grouping are opinion examination, extortion identification, and spam recognition and so forth Assessment mining is significantly being utilized for decisions, notice, business and so forth.

Verma et al. [8] dissected Sentiments of Indian government projects with the assistance of the vocabulary based word reference. The AI has adjusted the point of view of conclusion by giving incredible outcomes to infections like diabetes and epilepsy.

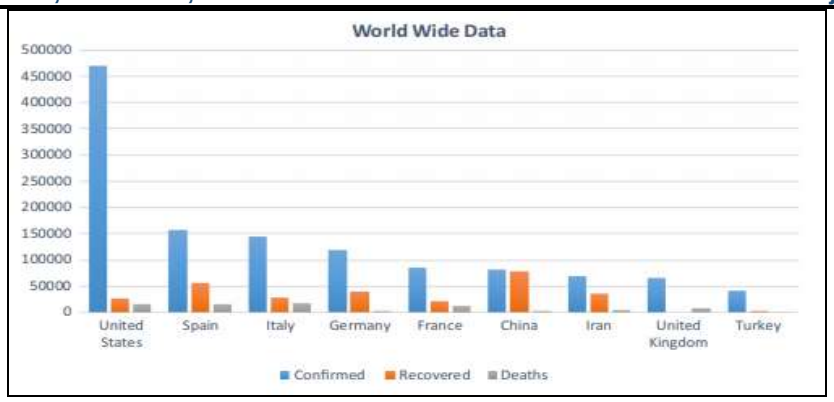


Fig. 1 Worldwide coronavirus as of 10th April 2020

Chakraborti et al. [9] distinguished epilepsy utilizing AI draws near, electroencephalogram (EEG) signals are utilized for recognizing ordinary and epileptic conditions utilizing fake neural organizations (ANN).

Sarwar et al. [10] conclusion diabetes utilizing AI and troupe learning methods result demonstrated that outfit strategy guaranteed precision of 98.60%. These reasons can be helpful to analyze and anticipate COVID-19. Firm and definite determination of COVID-19 can save a great many lives and can create a gigantic measure of information on which an AI (ML) models can be prepared. ML may give valuable contribution to this respect, specifically in making analyze dependent on clinical content, radiography Images and so forth.

As indicated by Bullock et al. [11], Machine learning and profound learning can supplant people by giving an exact determination. The ideal conclusion can save radiologists' time and can be savvy than standard tests for COVID-19. X-beams and registered tomography (CT) outputs can be utilized for preparing the AI model. A few drives are in progress in such 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 recognized 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 forecast who will be influenced all the more seriously can help in coordinating help and arranging clinical asset assignment and usage.

Yan et al. [13] utilized AI to foster a prognostic forecast calculation to anticipate the mortality hazard of an individual that has been tainted, utilizing information from (just) 29 patients at Tongji Hospital in Wuhan, China.

Jiang et al. [14] proposed an AI model that can foresee an individual influenced with COVID-19 and has the likelihood to foster intense respiratory pain disorder (ARDS). The proposed model came about in 80% of precision. The examples of 53 patients were utilized for preparing their model and are limited to two Chinese medical clinics. ML can be utilized to analyze COVID-19 which needs a ton of examination exertion yet isn't yet broadly operational. Since less work is being done on conclusion and anticipating utilizing text, we utilized AI and troupe learning models to characterize the clinical reports into four classifications of infections

3. EXISTING SYSTEM AND ITS LIMITATIONS

The infection may communicate between people through different roots which are right now hazy. The contaminated individual shows manifestations inside 2–14 days, contingent upon the hatching time of the center east respiratory condition (MERS), and the serious intense respiratory disorder (SARS). As indicated by W.H.O the signs and side effects of gentle to direct cases are dry hack, weakness and fever while as in extreme cases dyspnea (windedness), Fever and sluggishness may happen [3, 4]. The people having different sicknesses like asthma, diabetes, and coronary illness are more powerless against the infection and may turn out to be seriously sick. The individual is analyzing dependent on manifestations and his movement history. Crucial signs are being noticed definitely of the customer having manifestations. No particular treatment has been found as on tenth April 2020, and patients are being dealt with apparently.

The medications like Hydroxychloroquine, antipyretic, hostile to virals are utilized for the indicative treatment. At present, no such immunization is produced for forestalling this lethal infection, and we may play it safe to forestall this illness. By washing hands routinely with cleanser for 20 s and staying away from close contact with others by keeping the distance of around 1 m may diminish the odds of getting influenced by this infection. While wheezing, Covering the mouth and nose with the assistance of dispensable tissue and keeping away from the contact with the nose, ear and mouth can help in its avoidance. SARS is an airborne infection that showed up in 2003 in China and influenced 26 nations by having 8 K cases around the same time and moved from one individual to another

4. PROPOSED SYSTEM AND ITS ADVANTAGES

Machine learning provides a lot of support in identifying the disease with the help of image and textual data. Machine learning can be used for the identification of novel coronavirus. It can also forecast the nature of the virus across the globe. However, machine learning requires a huge amount of data for classifying or predicting diseases. Supervised machine learning algorithms need annotated data for classifying the text or image into different categories. From the past decade, a huge amount of progress is being made in this area for resolving some critical projects. Recent pandemic has attracted many researchers around the globe to solve this problem. Data provided by John Hopkins University in the form of X-ray images and various researchers build a model of machine learning that classifies X-ray image into COVID-19 or not. Since the latest data published by Johns Hopkins gives the metadata of these images. The data consists of clinical reports in the form of text in this paper, we are classifying that text into four different categories of diseases such that it can help in detecting coronavirus from earlier clinical symptoms. We used supervised machine learning techniques for classifying the text into four different categories COVID, SARS, ARDS and Both (COVID, ARDS). We are also using ensemble learning techniques for classification.

5. PROPOSED METHODOLOGY

The proposed methodology consists of 5.1 to 5.5 steps. In step 5.1 data collection is being performed and 5.2 define the refining of data, 5.3 gives an overview of preprocessing, 5.4 provides a mechanism for feature extraction. In E traditional machine learning algorithms are discussed and 5.5 give an overview of ensemble machine learning algorithms. The visual representation of the proposed methodology is shown in Fig. 2. and are being discussed below.

5.1 DATA COLLECTION

As W.H.O pronounced Coronavirus pandemic as Health Emergency. The analysts and medical clinics give open admittance to the information in regards to this pandemic. We have gathered from an open-source information vault GitHub.1 In which around 212 patients information is put away which have shown indications of Covid and other infections. Information comprises of around 24 ascribes to be specific patient id, counterbalance, sex, age, discovering, endurance, intubated, went_icu, needed_supplemental_O2, intubated, temperature, pO2_saturation, leukocyte_count, neutrophil check, lymphocyte tally, see, methodology, date, area, envelope, filename, DOI, URL. Permit. Clinical notes and different notes.

5.2 RELEVANT DATASET

Since our work is with respect to message mining so we removed clinical notes and discoveries. Clinical notes comprise of text while as the quality discovering comprise name of the relating text. Around 212 reports were utilized and their length was determined. We consider just those reports that are written in the English language. Figure 3 gives the length circulation of clinical reports that are written in English. The clinical reports are named to their relating classes. In our dataset, we have four classes COVID, ARDS, SARS and Both (COVID, ARDS). Figure 4 shows the various classes wherein clinical content is being sorted and relating report length.

5.3 PREPROCESSING

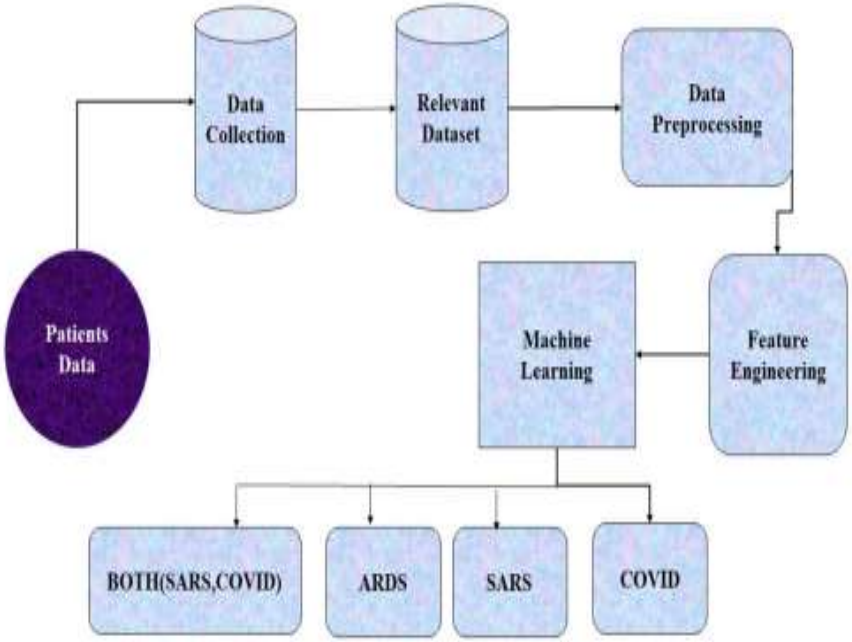
The text is unstructured so it needed to be refined such that machine learning can be done. Various steps are being followed in this phase; the text is being cleaned by removing unnecessary text. Punctuation and lemmatization are being done such that the data is refined in a better way. Stopwords, symbols, Url's, links are removed such that classification can be achieved with better accuracy. Figure 5 shows the main steps in preprocessing.

5.4 FEATURE ENGINEERING

From the preprocessed clinical reports, various features are extracted as per the semantics and are converted into probabilistic values. We use TF//IDF technique for extracting relevant features. Bag of words was also taken into consideration, unigrams, bigrams were also extracted. We identified 40 relevant features by which

the classification can be achieved. These features are shown in Fig. 6. By giving the corresponding weight to the feature and the same input is being supplied to machine learning algorithms.

Fig. 2 Methodology



Report length distribution

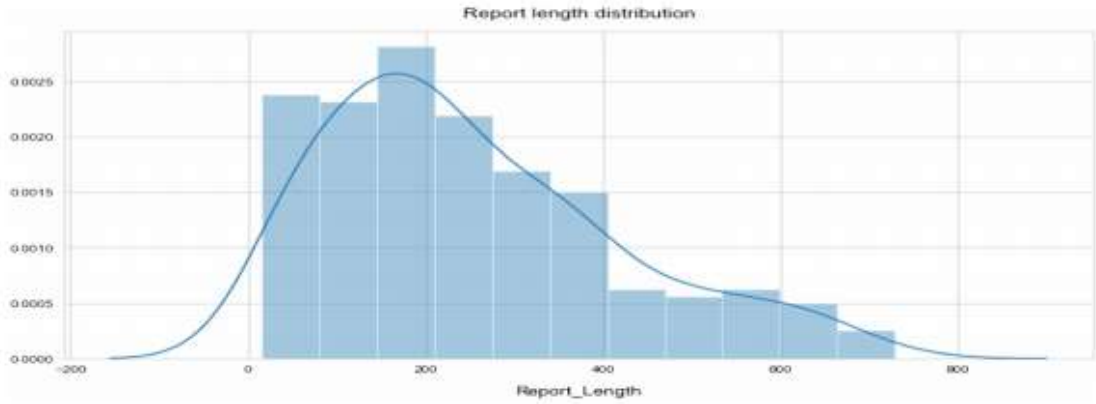


Fig. 3 Clinical report length

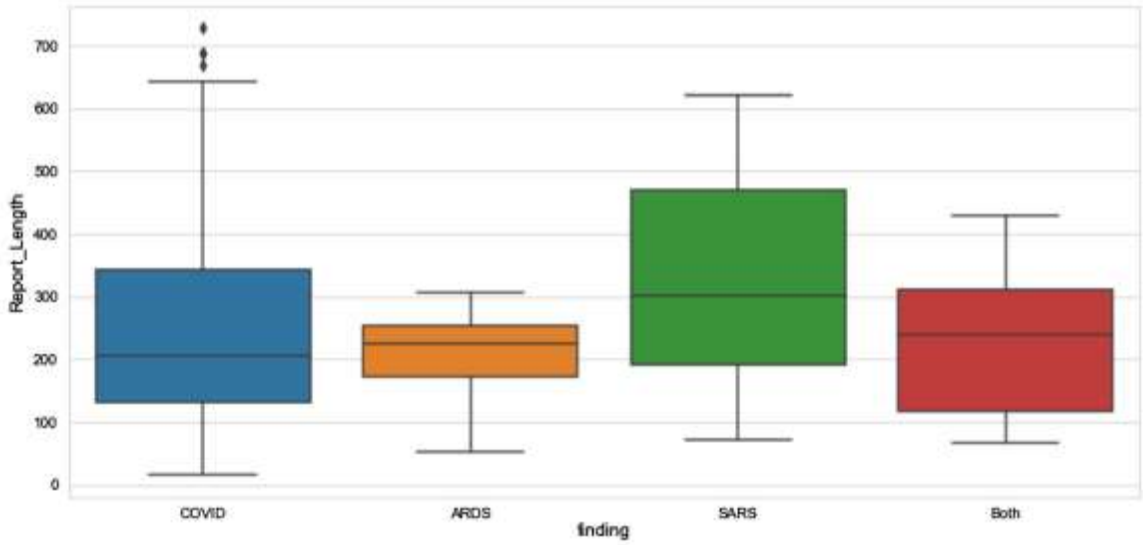


Fig. 4 Different classed with their report length

Clinical_notes	Finding	Report_Length	Punctuation	Lemmatisation	Stop_Word Removal
infiltrate in the upper lobe	COVID	45	infiltrate in the upper	infiltrate in the upper lobe	infiltrate upper lobe leave lung
progressive infiltrate and	COVID	40	progressive infiltrate a	progressive infiltrate and c	progressive infiltrate consolidation
progressive infiltrate and	COVID	40	progressive infiltrate a	progressive infiltrate and c	progressive infiltrate consolidation
progressive infiltrate and	COVID	40	progressive infiltrate a	progressive infiltrate and c	progressive infiltrate consolidation
diffuse infiltrates in the bil	COVID	48	diffuse infiltrates in th	diffuse infiltrate in the bila	diffuse infiltrate bilateral lower lungs
progressive diffuse inters	COVID	115	progressive diffuse int	progressive diffuse interst	progressive diffuse interstitial opacities consolidation
Severe ARDS. Person is in	ARDS	53	severe ards person is	severe ards person be intu	severe ards person intubate og place
Case 2: chest x-ray obtain	COVID	563	case 2 chest x-ray obt	case 2 chest x-ray obtain o	case 2 chest x-ray obtain jan 6 (2a) brightness lungs
Case 2: chest x-ray obtain	COVID	563	case 2 chest x-ray obt	case 2 chest x-ray obtain o	case 2 chest x-ray obtain jan 6 (2a) brightness lungs
SARS in a 74-year-old mar	SARS	71	sars in a 74-year-old r	sars in a 74-year-old man	sars 74-year-old man develop symptoms 4 days exp
SARS in a 74-year-old mar	SARS	71	sars in a 74-year-old r	sars in a 74-year-old man	sars 74-year-old man develop symptoms 4 days exp
SARS in a 74-year-old mar	SARS	71	sars in a 74-year-old r	sars in a 74-year-old man	sars 74-year-old man develop symptoms 4 days exp
SARS in a 29-year-old wor	SARS	378	sars in a 29-year-old v	sars in a 29-year-old wom	sars 29-year-old woman present 7 days exposure ()
SARS in a 29-year-old wor	SARS	378	sars in a 29-year-old v	sars in a 29-year-old wom	sars 29-year-old woman present 7 days exposure ()
SARS in a 42-year-old wor	SARS	145	sars in a 42-year-old v	sars in a 42-year-old wom	sars 42-year-old woman present 9 days exposure pc

5.5 MACHINE LEARNING CLASSIFICATION

The classification is performed to group the given content into four distinct sorts of infections. The four classes of infections, COVID (an individual having Covid), ARDS, SARS and both (comprises an individual that is having both Covid just as ARDS). Different directed AI calculations are being utilized to order the content into these classes. The AI calculations like help vector machine (SVM), multinomial Naı¨ve Bayes (MNB), strategic relapse, choice tree, irregular woodland, stowing, Adaboost and stochastic slope boosting were utilized for playing out this errand.

6. EXPERIMENTAL RESULTS

Implementation is a stage where the theoretical design is converted into a programmatic manner. In this proposed application we try to use PYTHON as a programming language in which Tkinter or Jupiter Notebook as a working platform to process the current application.

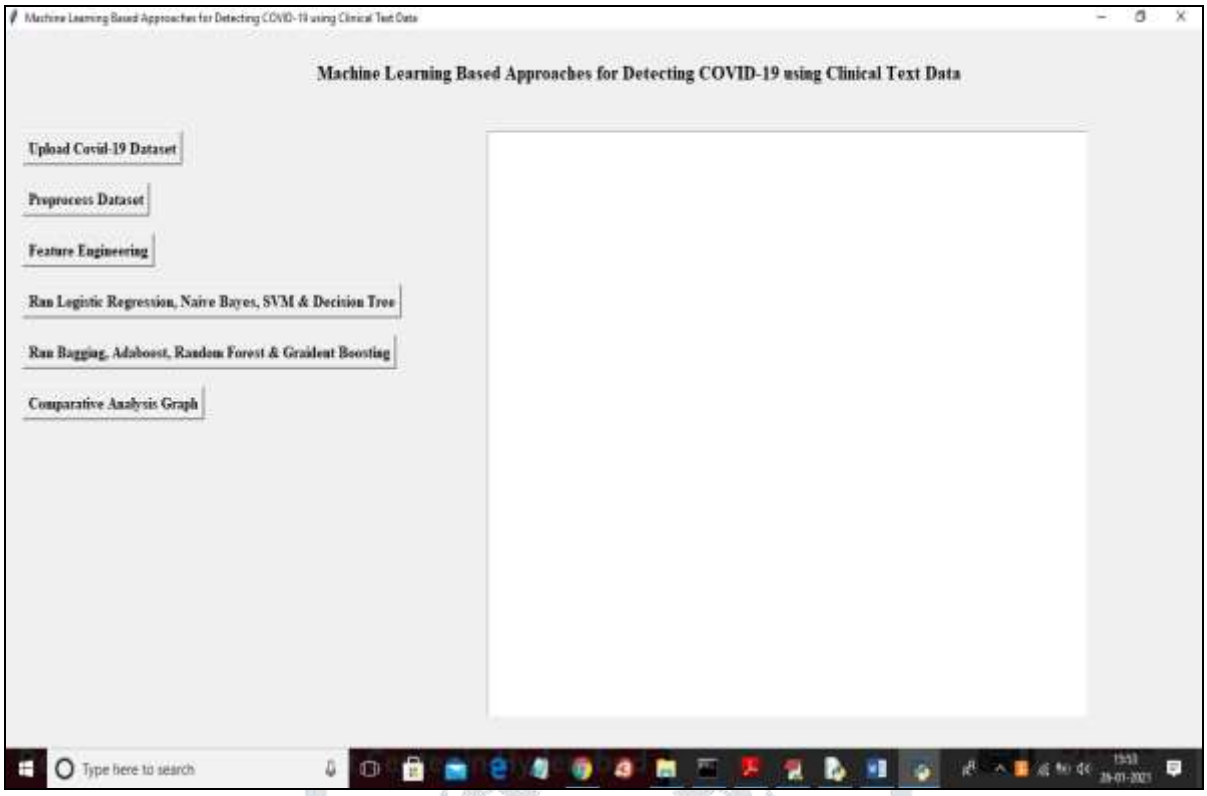
We utilized a windows framework with 4 GB Ram and 2.3 GHz processors for playing out this work. Sciket learn instrument is being utilized for performing AI grouping with the assistance of different libraries like NLTK, STOPWORDS and so on for improving the exactness of all the AI calculations pipeline is being utilized. In the

wake of playing out the factual calculation, more profound experiences about the information were 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 212 patients that are named into four classes. The grouping was finished utilizing AI calculations by providing them includes that were removed in the element designing advance. To investigate the speculation of our model from preparing information to inconspicuous information and lessen the chance of over fitting, 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 autonomously to keep away from the examining inclination presented by arbitrarily dividing the dataset in the cross-approval.

Table 1 gives a relative investigation of all the traditional AI strategies that are utilized for playing out this undertaking. Table 2 gives a near investigation of all the old style AI and Ensemble learning techniques that are utilized for playing out the assignment of ordering the clinical content into four classes. The outcomes showed that calculated relapse and Multinomial Nai¨ve Bayes Algorithm shows preferable outcome over any remaining calculations by having exactness 94%, review 96%, F1 score 95% and precision 96.2% different calculations like arbitrary woods, slope boosting likewise showed great outcomes by having exactness 94.3% separately. The envisioned near examination of the relative multitude of calculations that are utilized in our work is appeared in Fig. 7. Since we as a whole know, the COVID-19 information is least accessible.

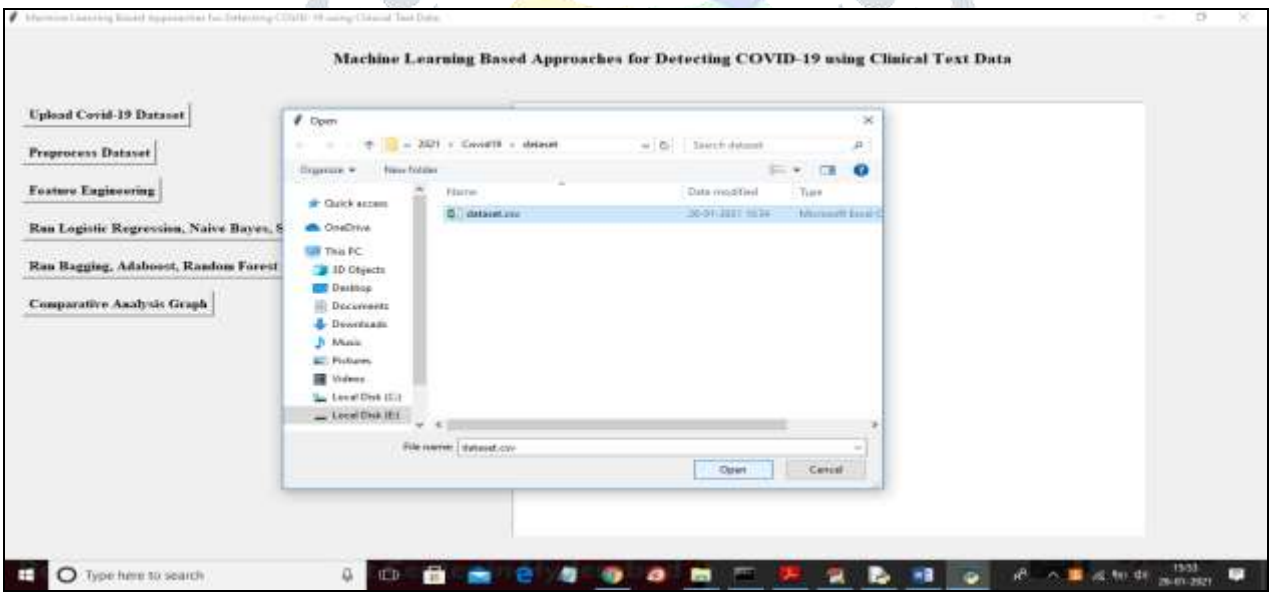
To get the genuine exactness of the model we tested it in two phases. In the principal stage, we took 75% of the accessible information and it shows less precision when contrasted with the stage where entire information was utilized for experimentation. So we can reason that if more information is provided 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 some way or another assistance the local area by dissecting the clinical reports and make important moves. Additionally, it was dissected that the COVID-19 patients report length is a lot more modest than different classes and it goes from 125 characters to 350 character

MAIN WINDOW



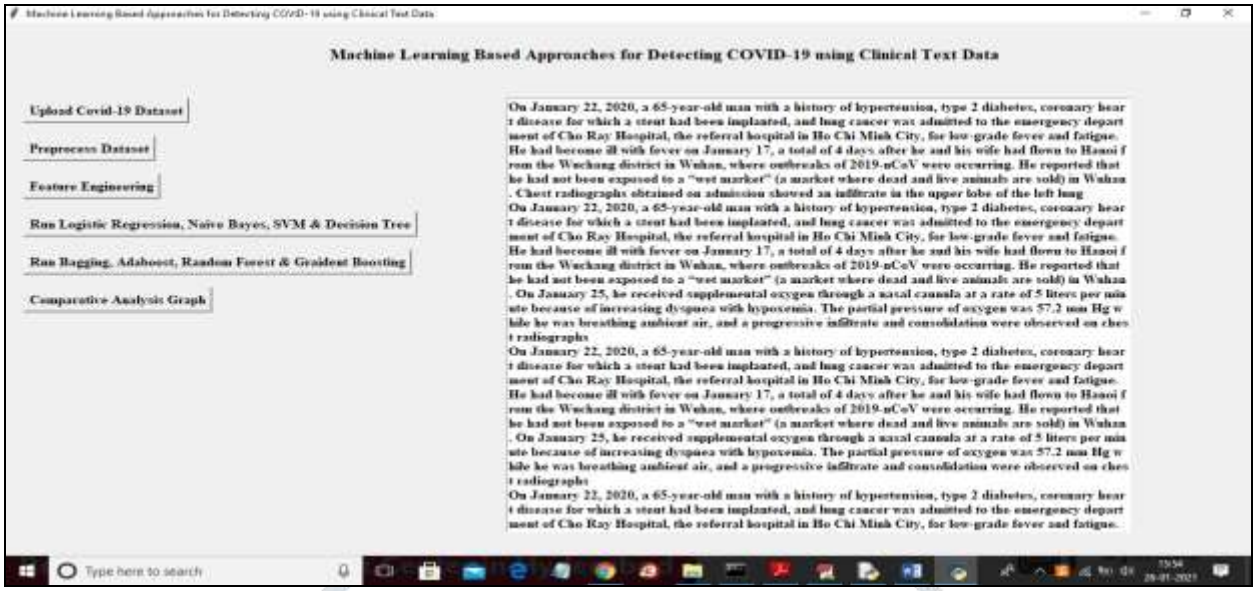
In above screen click on 'Upload Covid-19 Dataset' button and then upload dataset

LOAD THE DATASET



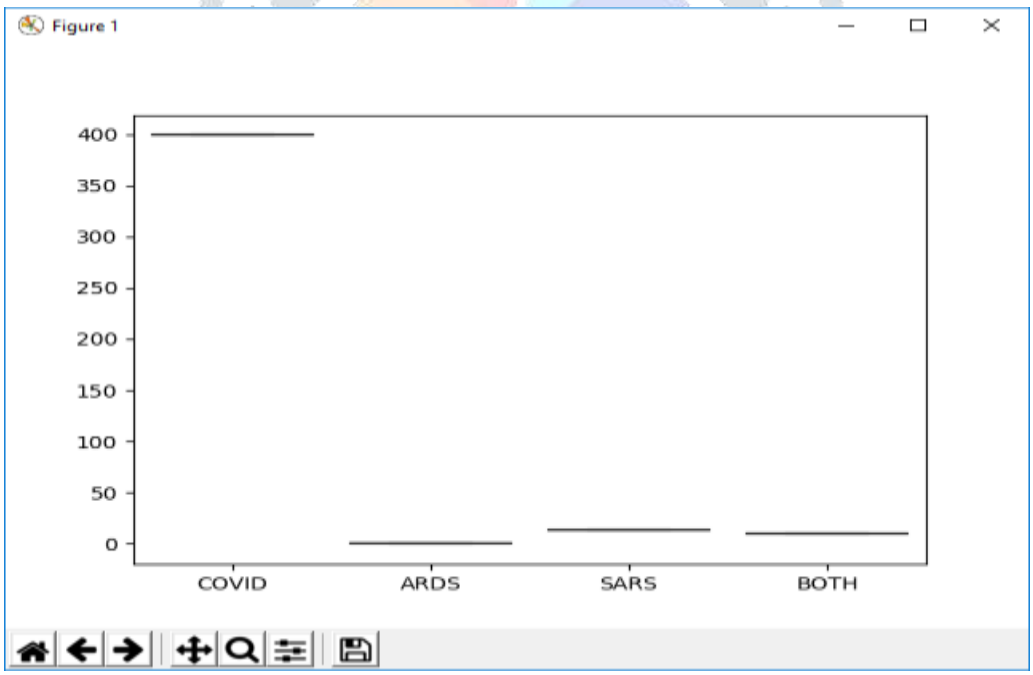
In above screen selecting and uploading 'dataset.csv' file and then click on 'Open' button to load dataset and to get below screen

DATA PRE-PROCESSING WINDOW



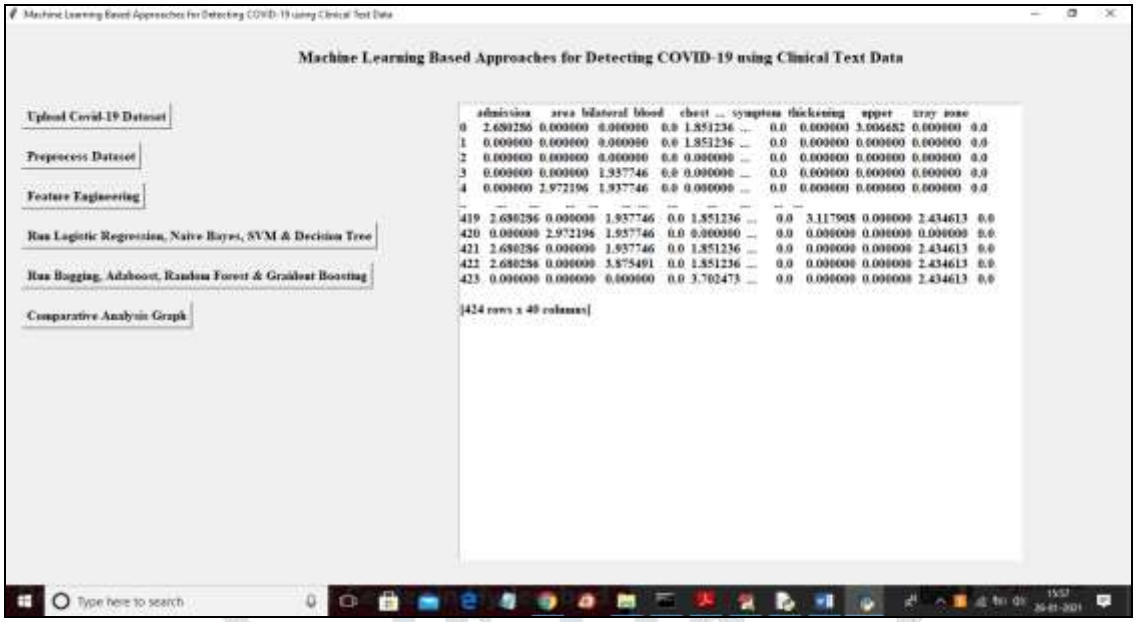
In above screen we extract all text data from dataset and now in above screen text in first sentence we have 'on' stop words and many number of numerical values and to remove those stop words and to clean data then click on 'Preprocess Dataset' button

FEATURE ENGINEERING WINDOW

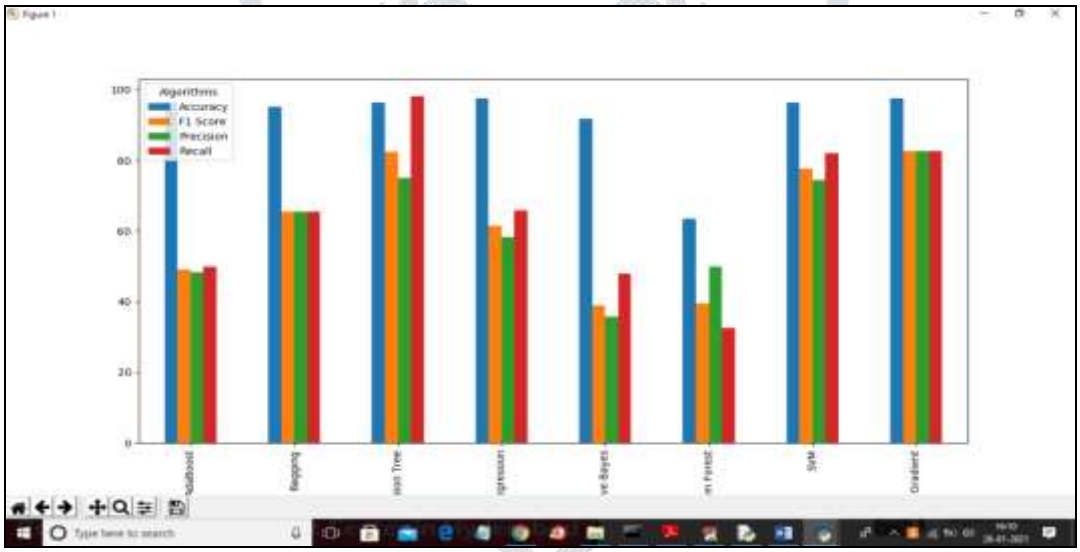


In above graph showing count/finding of each label and now close above graph and then click on 'Feature Engineering' button to apply TF-IDF on above text data and to get below features

APPLY SEVERAL ALGORITHMS TO CALCULATE ACCURACY



PERFORMANCE ANALYSIS WINDOW



In above screen we can see accuracy, precision, recall and fscore for each algorithm in group bar chart and in above graph x-axis represents algorithm name y-axis represents values

7. CONCLUSION

Coronavirus has stunned the world because of its non-accessibility of antibody or medication. Different scientists are working for vanquishing this lethal infection. We utilized 212 clinical reports which are marked in four classes to be specific COVID, SARS, ARDS and both (COVID, ARDS). Different highlights like TF/IDF, sack of words are being extricated from these clinical reports. The AI calculations are utilized for characterizing clinical reports into four unique classes. In the wake of performing characterization, it was uncovered that

calculated relapse and multinomial Naı¨ve Bayesian classifier gives brilliant outcomes by having 94% exactness, 96% review, 95% f1 score and precision 96.2%. Different other AI calculations that showed better outcomes were irregular timberland, stochastic angle boosting, choice trees and boosting. The effectiveness of models can be improved by expanding the measure of information. Additionally, the sickness can be characterized on the sexual orientation based with the end goal that we can get.

8. REFERENCES

1. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ (2020) A new coronavirus associated with human respiratory disease in china. Nature 44(59):265–269
2. Medscape Medical News, The WHO declares public health emergency for novel coronavirus (2020) https://www.medscape. com/viewarticle/924596
3. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Qiu Y, Wang J, Liu Y, Wei Y, Xia J, Yu T, Zhang X, Zhang L (2020) Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 395(10223):507–513
4. World health organization: https://www.who.int/new-room/g-adetail/q-a-corronaviruses#:/text=symptoms. Accessed 10 Apr 2020 5. Wikipedia coronavirus Pandemic data: https://en.m.wikipedia. org/wiki/Template:2019%E2%80%9320_coronavirus_pan demic_data. Accessed 10 Apr 2020
5. Khanday, A.M.U.D., Amin, A., Manzoor, I., & Bashir, R., ‘‘Face Recognition Techniques: A Critical Review’’ 2018
6. Kumar A, Dabas V, Hooda P (2018) Text classification algorithms for mining unstructured data: a SWOT analysis. Int J Inf Technol. https://doi.org/10.1007/s41870-017-0072-1
7. Verma P, Khanday AMUD, Rabani ST, Mir MH, Jamwal S (2019) Twitter Sentiment Analysis on Indian Government Project using R. Int J Recent Tech Eng. https://doi.org/10.35940/ijrte. C6612.098319
8. Chakraborti S, Choudhary A, Singh A et al (2018) A machine learning based method to detect epilepsy. Int J Inf Technol 10:257–263. https://doi.org/
9. Sarwar A, Ali M, Manhas J et al (2018) Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. Int J Inf Technol. https://doi.org/10.1007/s41870-018-0270-5
10. Bullock J, Luccioni A, Pham KH, Lam CSN, Luengo-Oroz M (2020) Mapping the landscape of artificial intelligence applications against COVID-19. https://arxiv.org/abs/2003.11336v1
11. Wang L, Wong A (2020) COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 Cases from chest radiography images. https://arxiv.org/abs/2003.09871