International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1081-1086 www.ijaem.net ISSN: 2395-5252

Exigency of Natural Language processing to skirmish with COVID-19 pandemic

Nidhi Agarwal[1], Manjeet Kaur Ratan[2]

Nidhi Agarwal Research Scholar (Dept of CSE) Indira Gandhi Delhi Technical University for Women, Delhi (India)*

Manjeet Kaur Ratan Software Associate IINCORE SOFTWARE SYSTEM Noida(India)

Submitted: 25-06-2021 Revised: 04-07-2021 Accepted: 07-07-2021

ABSTRACT

The COVID-19 pandemic outburst has affected whole world population. It needs urgent attention because of non-availability of properly tested vaccine and alarmingly fast spreading nature, ultimately effecting global medical management system. Besides spotting possible outbreaks as early as possible, a prime attention area in such a pathetic global situation is to boost a layman's confidence in analysing pre-treatment concerns and judging whether he is actually a patient or not. All this needs to be done in real time. NLP can be the front-runner in assessing and improving the quality of healthcare system by analysing existing patients' data (though less), opening up opportunities to make them more aware of their health and thus letting individuals take informed preliminary decisions on their own. Scraping the COVID-19 disease symptom webpages using combinative NLP tools and ML algorithms is a desirous and novel approach.

Keywords: Covid-19, Artificial Intelligence, Natural Language Processing, Pre-treatment, Pandemic

I. INTRODUCTION

This fresh coronavirus was detected first in Wuhan city of China from a sea food market. Since then, it has been declared as a pandemic because of its global threat [1]. The virus transmits very fast from one person to another, mainly by touching or by small nasal or fluid droplets in air. According to one of the strong predictions by a team of pandemic virus experts, this newly discovered coronavirus is presumed to spread for next 1.5 to 2 more years till it infects 65% to 75% of the world population [2]. Being very new in nature, a lot of studies and researches are going on in the whole world to get appropriate vaccine as early as possible to prevent global population from its fatal nature.

Blending artificial intelligence with natural language processing tools can be very desirous and novel in enabling even a layman make more aware of his symptoms to let him take pre-treatment decisions on his own [3-4]. Also, as patient data is increasing tremendously day by day, people being rejected for beds in hospitals, self-take care is most advisable in such a scenario. Rushing to hospitals is needed in severe cases only. The fatality rate is around 3% in the world as on 11 Nov 2020 [5-6]. According to the data of today (11.11.2020), total count of this virus positive cases globally is 52,422,158, out of which 36,668,563have recovered and 1,288,900have died. So, the total number of active cases are 14,464,695 [6]. The top 10 countries, as of today, according to the total count of infected persons are USA, India, Brazil, France, Russia, Spain, Argentina, UK, Colombia and Italy[6]. Figure 1 shows total count of persons infected, recovered and dead in top 5 countries since the onset of fatal virus (data as ontoday 11.10.2020).

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1081-1086 www.ijaem.net ISSN: 2395-5252

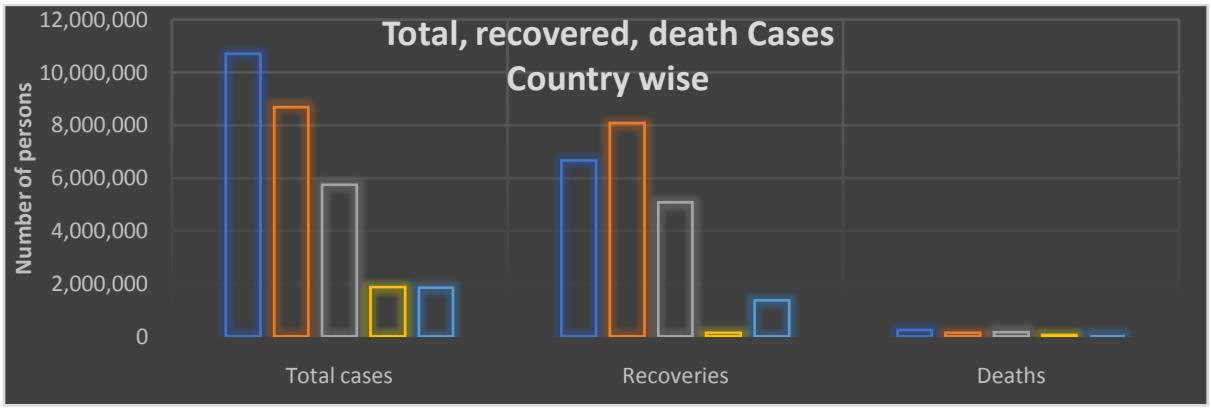


Fig. 1 Country wise data of total, recovered, dead persons

Figure 2 depicts the number of active cases in above 5 countries on the last day of every month from June to October [7] and Table 1 shows its data.

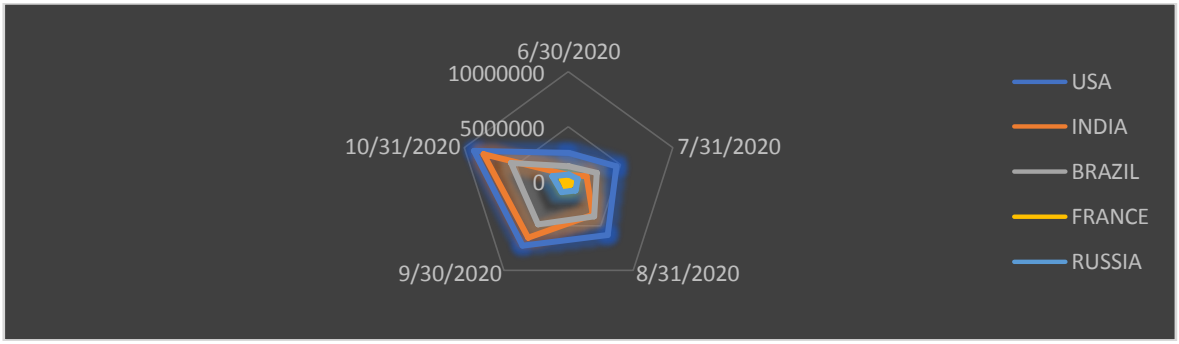


Fig. 2 Number of active COVID cases country-wise date-wise

	USA	INDIA	BRAZIL	FRANCE	RUSSIA
30-06-2020	**2628321**	**585481**	**1402041**	**164801**	**646929**
31-07-2020	**4549671**	**1695988**	**2662485**	**187919**	**838461**
31-08-2020	**6012569**	**3691166**	**3908272**	**281025**	**992402**
30-09-2020	**7213255**	**6312584**	**4810935**	**563535**	**1170799**
31-10-2020	**9122666**	**8184082**	**5535605**	**671638**	**1606267**

Table 1Number of active COVID cases country-wise date-wise

Though a lot of statistical data regarding number of cases active, dead and recovered is available for all COVID effected countries. Many Governmental and Private Bodies are making financial and food contributions, Government is declaring relief packages for COVID casualties on duties.[7]. But still the layman is unaware of what to do, where to go, whether to go or not, whether report to anyone or not when initially facing corona like symptoms. There is no tool or technology available which can assist a common man in the various countries effected by COVID-19. One cannot deny the fact that we have very fewer patient data available with us which can be trained. But scarce data can even assist, guide and perhaps save the lives of many people if scraped and summarized properly using advanced NLP tools. There is an urge for development of tool based on Natural Language Processing working in real time to read all the available data, symptoms, intensity of attack based on all possible patients' data. Although countries' rules do not allow making patient history public, but since this new virus has changed even the birth, marriage, death rituals, so these rules can also be forfeited until proper vaccine is developed. As far as lesser available data problem is concerned, it can be overcome gradually by adding new outcomes as a part of database only. Principal focus in this paper is towards drawing attention of concerned authorities for the requirement of an official website containing all patients' data with all possible attributes(updated

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1081-1086 www.ijaem.net ISSN: 2395-5252

after some fixed time interval) letting strong ML algorithms to scrap the database thus helping even a layman confidently follow the pre-treatment phase(decisively come out of his house, if needed) following the experiences and matching the disease intensities from the dataset. As cases are growing very fast all over the world and medical facilities are limited, coping up with each individual's needs becomes tough. Such a tool would also be quite helpful for pre-assessment and pre-diagnosis scenarios like confidence building for home quarantine, self-curing of mild symptoms.

II. INFERENCE FROM PANDEMIC DATABASE USING NLP APPROACH

The proposed framework will investigate, analyse and develop a tool to help compare a particular individual the intensity of his corona intensity symptoms on the basis of patient history already available on the proposed Governmental website. The patients' database website can be inferred using the following steps of applications of Artificial Intelligence as depicted by Figure 3 below.

Data Extraction using NLP
Data Evaluation using k-means clustering
Information prediction using Naïve Bayes classification.

Fig 3 Steps of the proposed approach

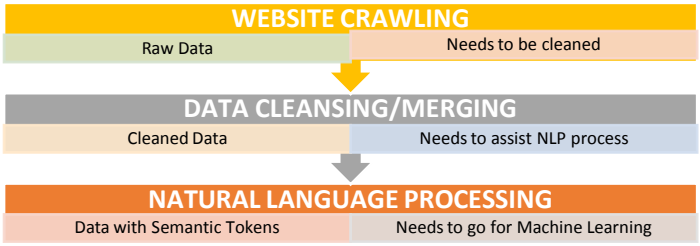
First of all, the relevant patient data will be extracted from the proposed Governmental patient website using Web Scraping method in the tabular form with various involved attributes. Then, based on this tabular representation of data, various clusters would be identified using clustering process. Further, to identify the probability of sustaining in a particular cluster with underlying symptoms, Naïve Bayes classifier will be used. An example of a cluster named Home Quarantine is mentioned in Section 2.3 to ease the understanding of how algorithm works to help individual take pre-treatment decision son his own.

2.1 Data Extraction from patients' website

"Web" scraping is being adopted extensively since last many years as an efficient recipe used by data extraction tools. It starts with first segmenting the web pages followed by extracting pertinent knowledge. All the important information like patient age, blood group, medical history, abroad visits, intensity of COVID-19 can be procured from the process of "web" scraping relying mainly on Process Automation tools for machine learning [8]. But the data obtained here is raw one containing various repeated ambiguous data which can further be grouped based on various attributes as per surfing patient's needs. All this is accomplished through data cleansing and data merging tools. Now NLP can be applied using various NLTK tools in real time like capitalisation, stop words removal, tokenisation and stemming. To extract data and process component, development of data "web-crawling" engine is proposed which can automate the extraction of all patient data to enumerate symptoms specifically from the proposed Governmental website. The extraction process will convert unstructured data to the structured one in the tabular form with various attributes marked in columns. Figure 4a depicts data extraction process including mainly following 3 stages:

Website crawling for generation of underdone dataset containing symptoms with prevailing conditions.
Data cleansing or merging
Natural Language Processing (NLP)

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1081-1086 www.ijaem.net ISSN: 2395-5252



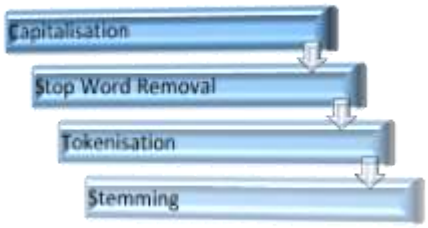


Fig 4a Data Extraction process steps**Fig 4b** NLP process

The raw data contains many repeating conditions based on various COVID symptoms. Thus, it needs to be cleaned up to further help in implementing NLP process effectively. Clean-up process deletes redundant underdone datasets by mingling common indications wrapped up under conditions' specifications.

NLP process is shown in Figure 4b and is accomplished by coding using Python's 'NLTK' library. The processing includes four main pre-processing methods

Capitalisation --→Stop word removal--→Tokenisation and Stemming.

Capitalization is a very essential step to pre-process the text as textual data mostly involves contains blend of capital words. For example, any piece of code in any programming language considers "Natural" and "natural" as completely different words, though possessing the similar meaning. One way to resolve this is to convert data completely to lowercase. Stop word removal will assist on getting rid of those words in data which are not meaningful. An example can be the filler words like 'of' or 'to' are unimportant and don't have any meanings. Their removal also helps in reducing dimensional space thus providing ease while dealing with collection of word methods like term-frequencies and inverse document's frequencies [9]. Tokenisation, being quite direct process, converts paragraphs to sentences and sentence to words. As now our database contains the indications of every COVID situation with corresponding patient data wrapped under paragraph, usage of NLTK tokenisation methodology subsequently will break paragraphs to tokens of sentences thus easing their usage to cluster and classify algorithms. Stemming process is used to remove the existing suffixes from any word and thus convert it further to corresponding root stem. It then eases in normalising the data and improving the standards of data dictionary. To accomplish it, usage of an inbuilt stemming algorithm called Snowball stemmer is suggested. Algorithms for all these operations are built in with the NLTK tool. We just need to feed the pre-processed data from the patients' web page and obtain desired results. The data obtained after this step will look as depicted by Table 2 below.

Patient ID	Patient Name	Age	Sex	Blood Group	Travel History	Medical History	Immunity Level	Other chronic Disease	Survived
1	AA	5	M	A+	Yes	Normal	Very low	No	No
2	BB	15	F	O+	Yes	Chronic	Low	Yes	Yes
3	CC	25	M	B-	No	Highly chronic	High	Yes	No
4	DD	65	F	O+	No	Normal	Average	No	Yes

Table 2 Data obtained after Data Extraction Process

All underlying important attributes will be displayed in the tabular representation of data. If some data is missing for a particular attribute, it will be procured from the hospital or by contacting the individual taking help of local authorities. We are already running with scarce data, can't compromise to lose it further even for a single attribute value.

2.2 Data Evaluation

The cleaned and merged data is ready to be evaluated on the basis of various symptoms, time of recovery, self-quarantine at home, hospital quarantine, recovery rate for both types of quarantine, intensity of case to identify self as patient, intensity for immediate medical help, precautions, home remedies in case of mild symptoms etc. All this will elevate confidence among new patients especially with mild

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1081-1086 www.ijaem.net ISSN: 2395-5252

symptoms. Also, this system will warn casual people for the intensity and dire consequences of the disease. Small clusters of data can be made based on attributes like age, travel history, medical history, immunity level etc. k-means clustering algorithm will be most suitable for this work [10-13]. Nearly 5 clusters can be identified as under Self-Home Quarantine, Strict home quarantine,Strict home quarantine with all family members' isolation, Consult the concerned and Immediate Medical Attention

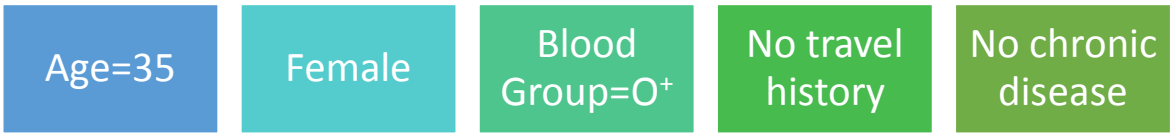


Fig. 5 Cluster1- Self Home quarantine

The likely patient will match symptoms and check in which cluster he is lying according to his prevailing medical conditions. The attributes will remain almost same in all the clusters, difference lies in their intensity values. In Figure 5 is shown an example of a cluster Self Home Quarantine with its underlying attributes.

2.3 Information Prediction

The prediction of the probability of falling in a particular cluster can be done using Naïve Based Classification [14] implemented in Python. Python language can be preferred as the NLTK tools for data scraping also use python programming. Figure 6 depicts COVID prediction process by analysing the combination of prevailing symptoms. NLP and Clustering processes are earlier discussed in above sections.

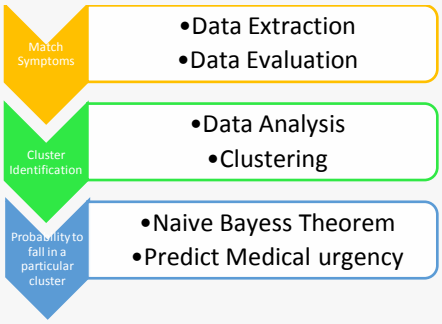


Fig. 6 Data Predication process

Here we emphasis mainly on the antipodal information prediction process adopting Naïve Bayes probability classifier algorithm. Firstly, prior probability of every cluster is calculated in the available patient dataset using the method beneath.

$$P(\text{cluster1}) = \frac{\text{count(cluster1)}}{\text{size(database)}}$$

It means that to calculate the probability value for any fresh input falling in the cluster1, preceding plausibility for that particular cluster is multiplied by the prospects of the input in the cluster.

Now we need to calculate the prospects of a symptom plunging in a particular cluster using the formula below.

$$P(\text{symptom}/\text{cluster1}) = \frac{\text{count(symptom, cluster1)} + 1}{\text{count(cluster)} + |N|}$$

It means to evaluate the plausibility of a particular disease ailment lying in a specific cluster, ration of total number of times the symptom fell in that cluster with Laplace smoothing 1 and the total number of different symptoms in that class. Now evaluate the necessary to find out the cluster for new unused data. This can be done using the following method [12].

$$P(\text{cluster1}/\text{input}) \sim P(\text{cluster1}) * P(\text{input}/\text{cluster1})$$

It means to calculate the value of probability for the new input falling in a cluster, calculate the product of preceding probability of the cluster with the probability for input within that cluster.

P(cluster1/mild_fever**age_38**sex_male**travelled_singapore_20_days** **_back**blood_group_A+ **Heart_patient**) ~ **P**(cluster1) * **P**(age_38/cluster1)* **P**(sex_male/cluster1)* **P**(mild_fever/cluster1)* **P**(travelled_singapore_20_days _back /cluster1)* **P**(blood_group_A+/cluster1)* **P**(Heart_patient/cluster1)

Same procedure can be repeated with every cluster to obtain resultant values to be used for information prediction results.

III. CONCLUSION

In this paper, we have proposed an approach of scraping the COVID-19 disease symptom webpages by adopting NLP tools available in NLTK library.All this is proposed to be done in real time to help even a layman take pre-treatment judgements confidently and efficiently. Artificial Intelligence can serve as a dominant and functional tool for tracking, pinpointing and monitoring infections caused by coronavirus COVID 19 among patients of differed severity levels. AI along with NLP can remarkably elevate treatment measures by implementing as well as by developing suitable algorithms. Combination of AI and NLP will further elevate the research for this completely new virus considering available patient dataset. AI with NLP can lend a hand to develop most suitable treatment tracks, anticipation methods, timely development of suitable vaccine.

Conflict of Interests The authors declare that they have no conflict of interest.

REFERENCES

[1]. Wu F, Zhao S, Yu B, Chen Y M, Wan g W, Song Z G, Hu Y, Tao Z W, Tian J H, Pei Y Y, Yuan M L, Zhang Y L, Dai F H, Liu Y, Wang Q M, Zheng J, Xu L, Holmes E C, Zhang Y Z (2020) A new coronavirus associated with human respiratory disease in china. Nature 44(59):265–269 https://doi.org/10. 1038/s41586-020-2008-3.

[2]. https://edition.cnn.com/2020/04/30/health/report-covid-two-more-years/index.html

[3]. Hu Z, Ge Q, Jin L, Xiong M (2020) Artificial intelligence forecasting of COVID-19 in China. arXiv preprint arXiv:2002.07112. 17 Feb 2020.

[4]. Haleem A, Javaid M Vaishya (2020) Effects of COVID 19 pandemic in daily life. CurrMed Res Pract.

[5]. WHO Report Coronavirus disease 2019 (COVID-19) Situation Report – 49 (2020 (accessed March 09 2020)) https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200229-sitrep-40-covid-19.pdf?sfvrsn=849d0665_2.

[6]. https://www.worldometers.info/coronavirus

[7]. https://mackuba.eu/corona/#united_states

[8]. Md Laskar T R, Md Hossain T A, Kamal R M Rashid N (2016) "Automated Disease Prediction System (ADPS): A User Input-based Reliable Architecture for Disease Prediction". International Journal of Computer Applications 133(15):24-29 January 2016. Published by Foundation of Computer Science (FCS) NY USA.

[9]. Aizawa A "An information-theoretic perspective of tf–idf measures". Information Processing and Management. 39 (1): 45–65. doi: 10. 1016/ S0306-4573(02)00021-3.

[10]. Coomans D, Massart D L "Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-Nearest neighbour classification by using alternative voting rules". Analytica Chimica Acta. 1982 136: 15–27. doi:10.1016/S0003-2670(01)95359-0.

[11]. Thomas J Princy R T (2016) "Human heart disease prediction system using data mining techniques " International Conference on Circuit Power and Computing Technologies (ICCPCT) Nagercoil 2016 pp. 1-5. doi: 10.1109/ICCPCT.2016.7530265

[12]. Tayeb S, Pirouz M, Sun J, Hall K, Chang A, Li J, Song C, Chauhan A, Ferra M, Sager T, Zhan J Latifi S (2017) "Toward predicting medical conditions using k-nearest neighbors". IEEE International Conference on Big Data (Big Data). [online] Available at: https://ieeexplore.ieee.org/document/8258395.

[13]. David M B, Andrew Y N, Michael I J (2003). Lafferty, John (ed.). "Latent Dirichlet Allocation". Journal of Machine Learning Research. 3 (4–5): pp. 993–1022. doi: 10. 1162/ jmlr. 2003. 3.4-5.993.

[14]. Pattekari S, Parveen A (2019) "Prediction System for Heart Disease Using Naïve Bayes". International Journal of Advanced Computer and Mathematical Sciences [online] 3(3) pp.290-294. Available at: https://pdfs.semanticscholar.org/ d32e/ e90a5de89093 a4fc95f43e0409cb91414726. pdf [Accessed 31 May 2019].