LETTER TO THE EDITOR

Predicting COVID-19 in Iran Using Automate Web Crawling

KIA JAHANBIN[1], VAHID RAHMANIAN*[2], FERESHTE RAHMANIAN[1]
[1]*PhD candidate in Information Technology, Yazd university, Yazd ,Iran*
[2]*MPH, Ph.D. in Epidemiology, Zoonoses Research Center, Jahrom University of Medical Sciences, Jahrom, Iran*
[3]*MSc of Information Technology, Islamic Azad University Branch of Kerman, Iran*
Correspondence to: Vahid Rahmanian: Zoonoses Email: rahmanian.vahid@ut.ac.ir , Tel: +989175985204
Research Center, Jahrom University of Medical Sciences, Jahrom, Iran,

Dear Editor

Collecting, analyzing and using data related to Covid-19 has become one of widely-debated research topics on this disease[1,2]. On the one hand, the use of social networks as a new method in identifying the prevalence of infectious diseases and predicting Covid-19 outbreaks has been proposed in various studies[1-3].

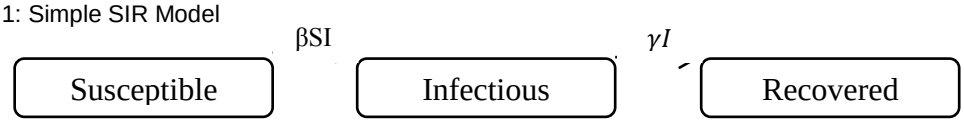
the subset sciences “Big Data” in social networking. A web news mining-based automatic system can monitor, evaluate, and categorize news, which, in addition to managing news articles, could also be applied to the field of advisory systems[4,5]

In a study titled a model called SIDARTHE, Giordano, G., et al. proposed that of the eight parameters susceptible (S), infected (I), diagnosed (D), ailing (A), recognized (R), threatened (T), healed (H) and (E) extinct have been used to predict the course of coronavirus disease in Italy[6].

In another study, using data obtained from the Ministry of Health during the period March 2, 2020 to May 15, 2020, two models are proposed to determine the end of the epidemic in Saudi Arabia using logistic growth and susceptible-infected-recovered[7].

In recent years, mathematical models are commonly used to predict the occurrence of the disease[8,9]. A mathematical model helps to predict a global or regional behavior of infection, or to draw an unreal world and change the protective laws of that future with very reasonable accuracy. One of the most common models used in epidemiology is the basic compartmental SIR model with Kermack and McKendrick, which is defined in a schematic view as follows[10](Fig.1):

Fig. 1: Simple SIR Model



Susceptible (S): Indicates the number of talented individuals. When a susceptible and infected person enters an "infectious contact", the infection is transmitted to the susceptible person and enters from the S chamber to the infectious chamber (I).

Infectious (I): Is the number of individuals infected, I's are people who are infected and are able to infect susceptible people.

Recovered (R): The number of individuals who are discharged (safe from illness) or dead. These are individuals who have become infected or have recovered from the disease and entered the chamber or died (it is assumed that the death toll is negligible given the total population), the chamber may also be "recovered" or "resistant".

As mentioned, in the SIR model individuals are divided into three groups: susceptible, infective and recovered. The functions S(t) and R(t) which indicate the behavior of the population in the time interval t (for example, t can indicate days) are defined by the following differential formulas:

$$\frac{dS}{dt} = -\beta SI$$

$$\frac{dI}{dt} = \beta SI - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

β (transfer rate): Is the average rate of infection transmission per time unit by making infected contact between an individual (I) and susceptible individual (in the S chamber).

ϒ (recovery rate): ϒ is the rate of recovery; ϒ / I the average length of time during which an infected person carries the disease and remains infected. Thus S (t) * I (t) β is the sum of the infection rates; in other words, it is the fraction of the population that is infected per time unit t.

The SIR model, also known as the SIR Basic, has other derivatives, as follows:

SIS model[11]: Some infections, such as colds and flu, do not provide long-term immunity, so in such infections, after recovery from the infection, immunity is not established and people return to the S-chamber (Fig.2):

Fig. 2: SIS model

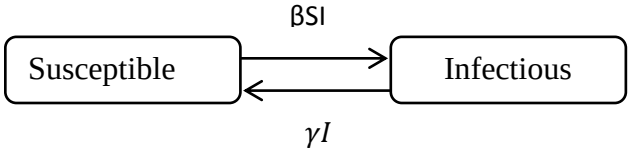


Fig.3: SIRD model

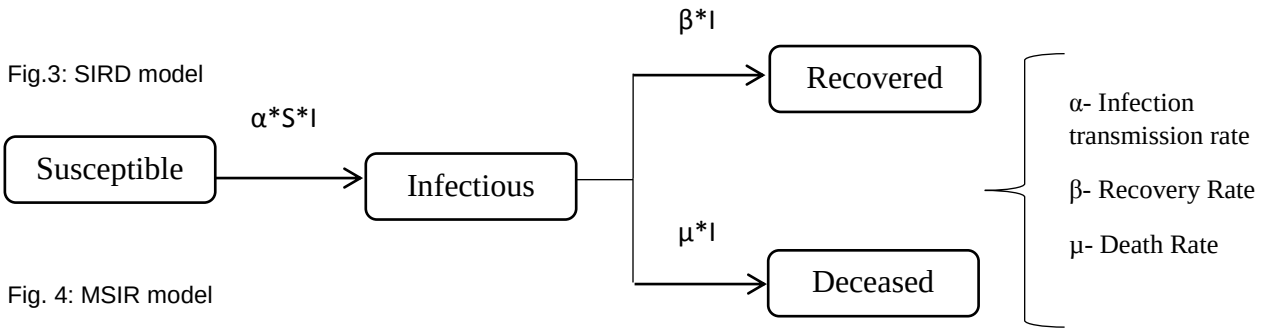


Fig. 4: MSIR model

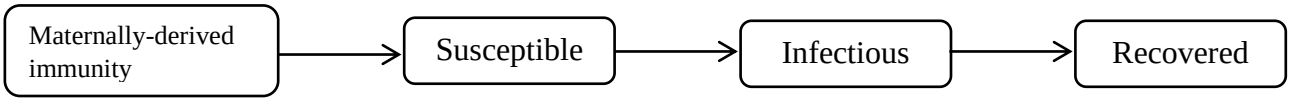
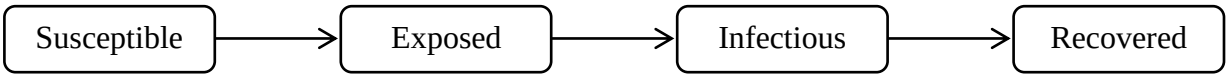


Fig. 5: SEIR model



1- SIRD model[12]: This model makes a special distinction between people who have survived the disease and are now immune and those who have died(Fig.3):
2- MSIR model[13]: In many infections, such as measles, infants are not born in an S-chamber due to protection by maternal antibodies (passing through the placenta and also through the colostrum) and are immune to the disease in the first few months of life. It is called passive immunity. This immunity (maternal-derived immune system) is shown by the M-chamber in the model(Fig.4):
3- The SEIR[14] model: In many infections, the latency period is very important. In this period, people are infected but are not yet aware of their infection. In this period, the person is exposed to the E-chamber (exposed) (Fig.5):
4- The SEIR[15] model: The SEIS model is similar to the SEIR model, except that the patient has no immunity after recovery.
5- In this study, the FAEMC-CD model, which has been shown to be effective in various studies[1,2], was first used. The implementation of this model is summarized as follows:

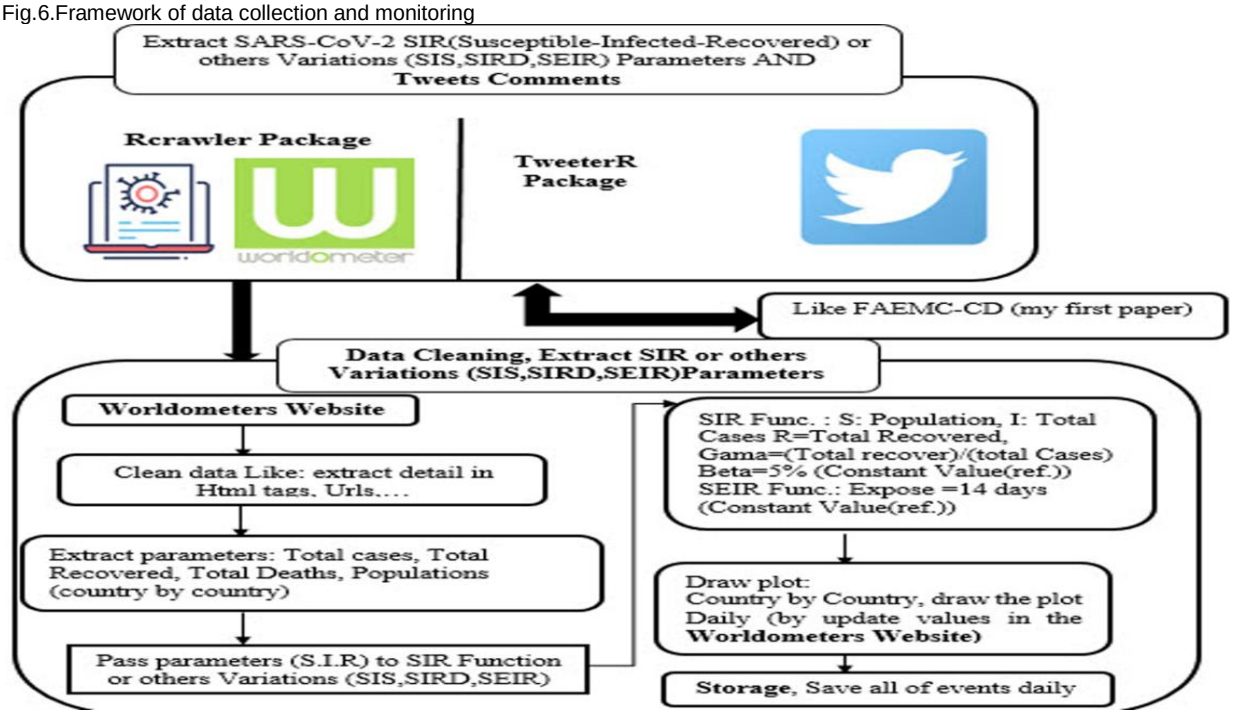
- Clear and merge data and extract vocabulary
- Web and tweet crawling
- Apply fuzzy rules and fuzzy classifiers

Data visualization: The first step is related to clearing, merging and integrating data and vocabulary extraction, which includes the steps: Unify letters (convert all letters to lowercase letters), tokenize, find stems, remove Slopworks (pronouns, auxiliary verbs, etc.) and term filtering with the TF-IDF method.

Evolving fuzzy rules were applied using the Fuzzy-Rule based Classification (FRBS) package[16]. The importance of the evolving fuzzy system in updating the terms extracted from the database is due to the possibility of adding new words or terms about diseases[17]; Finally, the visualized part of the proposed framework is used to monitor the moment of the outbreak of the disease, the speed of disease spread and take rapid and preventive measures to help decision-making centers in the field of health.

In this study, a new plugin was designed for the FAEMC-CD model. This enables the model to retrieve numbers by the Rcrawler R4.0[18] software package from websites that report up-to-date statistics and information on coronavirus disease (like Worldometers, Johns Hopkins University website and the World Health Organization), predict epidemic trends and curves based on SIR, SEIR, SIRD, and SIS models and provide them to researchers on a daily basis or during specific time periods. An overview of this plugin is shown in Fig. 6.

Fig.6.Framework of data collection and monitoring



The RCrawler package is able to extract the contents of web pages and generate data that can be used directly in other applications. The main features of RCrawler include multi-threaded crawling, duplicate content identification and extraction, URL filtering and content type, website search depth control and robot.txt parser. Using code (1) in R, the contents of the main table available on the Worldometers website can be extracted along with all the following links of the countries in the table:

Rcrawler(Website = 'https://www.worldometers.info/coronavirus/', no_cores = 4, no_conn = 4, ExtractXpathPat="//*[@id=main_table_countries_today]" MaxDepth = 3)

To determine the required Xpath from websites, to select the desired sections for extracting values, one might use different plugins. This study used the Xpath Generator plugin. Figure 7 shows the various Xpaths from the Worldometers website dated November 3, 2020, which show that the column for Total Cases has been selected and the corresponding Xpath has been retrieved.

Fig.7: The various Xpaths from the Worldometers website dated November 3, 2020,

#	Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	Active Cases	Serious, Critical	Tot Cases/ 1M pop	Deaths/ 1M pop	Total Tests	Tests/ 1M pop	Population
	World	47,706,820	+382,790	1,217,421	+6,349	34,225,093	12,264,306	87,958	6,120	156.2			
1	USA	9,616,165	+48,622	237,636	+640	6,188,778	3,189,751	17,597	28,994	716	150,266,637	453,069	331,663,944
2	India	8,312,947	+46,033	123,650	+511	7,654,757	534,540	8,944	6,004	89	111,789,350	80,737	1,384,604,543
3	Brazil	5,554,647	+441	160,282	+10	4,998,408	395,957	8,318	26,069	752	21,900,000	102,781	213,073,428
4	Russia	1,673,686	+18,648	28,828	+355	1,251,364	393,494	2,300	11,467	198	61,954,566	424,475	145,955,918
5	France	1,502,763	+36,330	38,289	+854	120,714	1,343,760	3,878	23,005	586	16,718,098	255,930	65,322,933
6	Spain	1,331,756	+18,669	36,495	+238	N/A	N/A	2,754	28,480	780	18,072,174	386,480	46,760,988
7	Argentina	1,183,131		31,623		998,016	153,492	4,922	26,097	698	3,047,313	67,215	45,336,503
8	Colombia	1,093,256		31,670		985,796	75,790	2,376	21,408	620	5,126,096	100,381	51,066,585
9	UK	1,073,882	+20,018	47,250	+397	N/A	N/A	1,075	15,791	695	34,634,393	509,272	68,007,595
10	Mexico	933,155	+3,763	92,100	+205	687,420	153,635	2,838	7,212	712	2,414,882	18,663	129,391,657
11	Peru	906,545		34,585		830,612	41,348	1,043	27,366	1,044	4,551,185	137,387	33,126,781
12	Italy	759,829	+28,244	39,412	+353	302,275	418,142	2,225	12,573	652	16,285,936	269,495	60,431,276
13	South Africa	728,836	+1,241	19,539	+74	659,259	50,038	546	12,237	328	4,868,610	81,741	59,561,343
14	Iran	637,712	+8,932	36,160	+422	495,473	106,079	5,378	7,560	429	5,036,633	59,707	84,356,102
15	Germany	575,196	+14,610	10,868	+134	371,500	192,828	2,388	6,858	130	21,882,967	260,898	83,875,505

After extracting the desired Xpath and applying to the ExtractXpathPat code section (1), all pages and sub-pages related to the table were extracted and after the data clearance process, the parameters required to send to the SIR algorithm or its derivatives were extracted from pages and subpages. The contents of the Worldometer's website, extracted by the RCrawler package, include 1865 Html pages (53MB). After deleting irrelevant pages, 1626 pages with a volume of 36.9 Mb were removed.

Finally, based on the extracted parameters related to the data reported by Covid-19 in Iran on the Worldometers website and considering the parameters

R0=2،

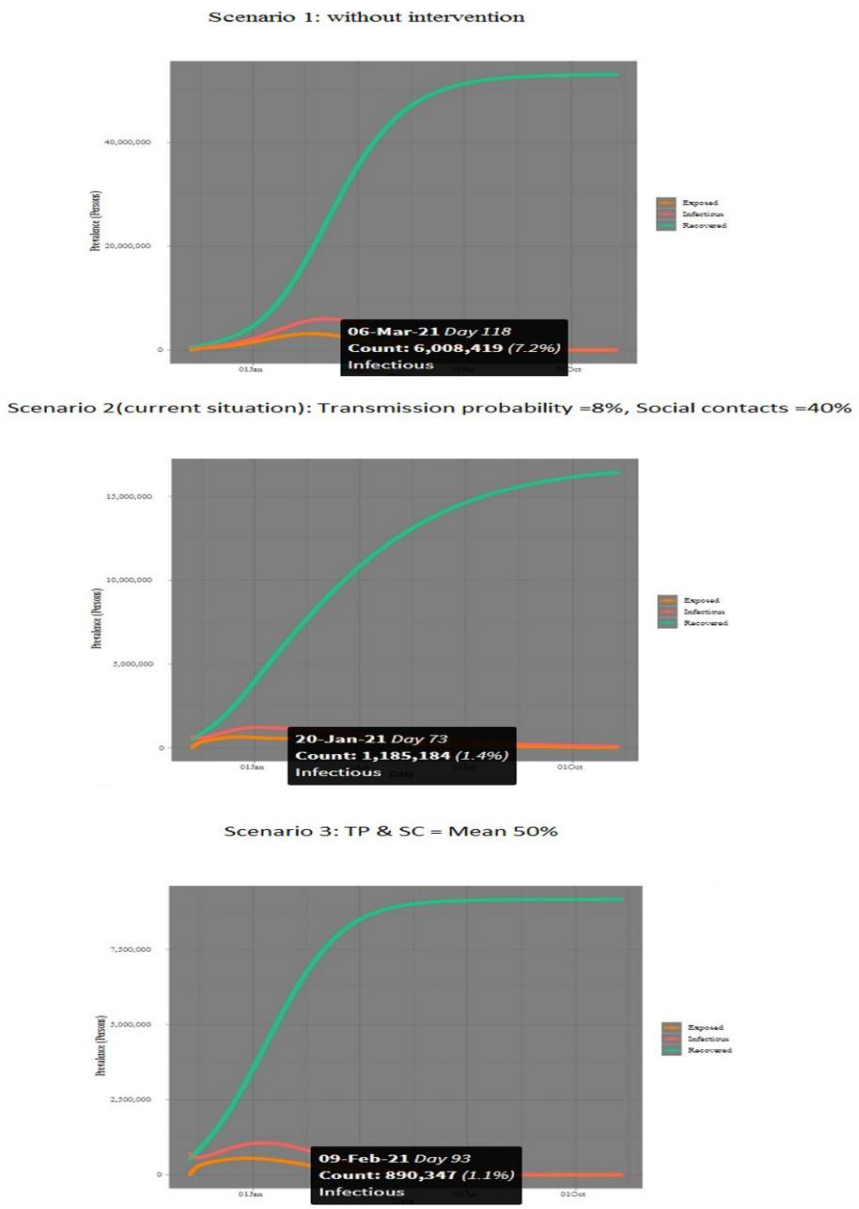
Duration of the latent period in days=7 days

Duration of infectious period in days=14 days

Number of days to Simulation=365

The trend of Covid-19 epidemic in Iran was predicted based on the SEIR model in different scenarios (Fig. 8).

Fig.8: Predicting the Covid-19 epidemic trend based on the SEIR model by extracting data from the Worldometer's website in three different scenarios,



Transmission probability - interventions that affect the likelihood of transmission of the virus between infected and susceptible individuals, like using a face mask and hand washing Social contacts: This intervention demonstrates the impact of actions that target the number of face-to-face contacts between people in the community, such as: Quarantine, school closures, housework and other "lockdown" activities.

The FAEMC-CD algorithm has been used in many previous studies in the infectious disease care system to monitor and predict epidemics. It has been shown that this model has high accuracy, classification error, kappa and absolute error. In the classification of documents related to experimental data[1,2].

In the disease surveillance system, it is very important to diagnose outbreaks and their process ahead of time. The use of epidemiological models with a mathematical background can lead health policy makers to this goal and provide an effective model for disease control and prevention[1,2,8,9].

Automatically extracting news and comments allows for accessing a big database of heat-treated events in a short period of time, saving time through using the reported data and automate different models.This study designed a new plugin for the FAEMC-CD model, which adds to the model the ability to automatically extract data from websites. The data retrieved from these websites are extracted by the plugin. After the data clearance process, it prepares the parameters needed to send to the algorithms such as SIR. Finally, it predicts the trend and draws epidemic curves based on them on a daily basis, helps policy makers and health planners for future interventions and appropriate evidence-based decisions over time.

In conclusion, the proposed model saves time and money based on updated data published on world-known websites to easily predict the future of the Covid-19 epidemic situation based on SIR models worldwide, countries or even provides states or provinces and makes better use of the data available on websites.

Conflict of interests: The authors declare that there is no conflict of interest

REFERENCES

1. Jahanbin K, Rahmanian F, Rahmanian V, Jahromi ASJGH, Control I. Application of Twitter and web news mining in infectious disease surveillance systems and prospects for public health. 2019; 14.
2. Jahanbin K, Rahmanian VJAPJoTM. Using Twitter and web news mining to predict COVID-19 outbreak. 2020; 13.
3. Depoux A, Martin S, Karafillakis E, Preet R, Wilder-Smith A, Larson H. The pandemic of social media panic travels faster than the COVID-19 outbreak. Oxford University Press; 2020.
4. Iglesias JA, Tiemblo A, Ledezma A, Sanchis AJIF. Web news mining in an evolving framework. 2016; 28: 90-8.
5. Škrjanc I, Iglesias JA, Sanchis A, Leite D, Lughofer E, Gomide FJIS. Evolving fuzzy and neuro-fuzzy approaches in clustering, regression, identification, and classification: a survey. 2019; 490: 344-68.
6. Giordano G, Blanchini F, Bruno R, Colaneri P, Di Filippo A, Di Matteo A, Colaneri MJNM. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. 2020: 1-6.
7. Alboaneen D, Pranggono B, Alshammari D, Alqahtani N, Alyaffer RJIJoER, Health P. Predicting the Epidemiological Outbreak of the Coronavirus Disease 2019 (COVID-19) in Saudi Arabia. 2020; 17: 4568.
8. Esmaeilzadeh N, Shakeri M, Esmaeilzadeh M, Rahmanian V. ARIMA models forecasting the SARS-COV-2 in the Islamic Republic of Iran. Asian Pacific Journal of Tropical Medicine 2020; 13: 521.
9. Rahmanian V, Bokaie S, Rahmanian K, Hosseini S, Firouzeh AT. Analysis of temporal trends of human brucellosis between 2013 and 2018 in Yazd Province, Iran to predict future trends in incidence: A time-series study using ARIMA model. Asian Pacific Journal of Tropical Medicine 2020; 13: 272.
10. Huppert A, Katriel GJCm, infection. Mathematical modelling and prediction in infectious disease epidemiology. 2013; 19: 999-1005.
11. Gray A, Greenhalgh D, Hu L, Mao X, Pan JJSJoAM. A stochastic differential equation SIS epidemic model. 2011; 71: 876-902.
12. Farboodi M, Jarosch G, Shimer R. Internal and external effects of social distancing in a pandemic: National Bureau of Economic Research2020. Report No.: 0898-2937.
13. Pinter G, Felde I, Mosavi A, Ghamisi P, Gloaguen RJM. COVID-19 Pandemic Prediction for Hungary; a Hybrid Machine Learning Approach. 2020; 8: 890.
14. Kai D, Goldstein G-P, Morgunov A, Nangalia V, Rotkirch AJapa. Universal masking is urgent in the covid-19 pandemic: Seir and agent based models, empirical validation, policy recommendations. 2020.
15. Dowdle WJBotWHO. Influenza A virus recycling revisited. 1999; 77: 820.
16. Riza LS, Bergmeir CN, Herrera F, Benítez Sánchez JM, editors. frbs: Fuzzy rule-based systems for classification and regression in R2015: American Statistical Association.
17. Angelov PP, Zhou X. Evolving fuzzy-rule-based classifiers from data streams. IEEE Transactions on Fuzzy Systems 2008; 16: 1462-75.
18. Khalil S, Fakir MJS. RCrawler: An R package for parallel web crawling and scraping. 2017; 6: 98-106.