JURNAL TEKNOINFO
Volume 16, Nomor 2, Juli 2022, Page 246-251
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

WEB SCRAPING OF DISEASE INFORMATION FROM SOCIAL MEDIA TWITTER

Muhammad Iqbal Habibie[1)], Taufiq Widiaputra[2)], Yulianingsani[3)]

[1,2,3]Agency for the Assessment and Application of Technology (BPPT)
[1,2,3]Jl. M.H Thamrin No. 8 Jakarta Pusat DKI Jakarta, Indonesia
Email: [1]muhammad.iqbal@bppt.go.id, [2]taufiq.widiaputra@bppt.go.id, [3]yulianingsani@bppt.go.id

Abstract

Environmental degradation caused by land conversion, trash (both domestic and industrial), and natural catastrophes is all variables that contribute to the establishment of disease susceptibility. Experts throughout the world suggest "ONE HEALTH" as a strategy for dealing with the threat of zoonoses. The One Health concept is a worldwide strategy to expand interdisciplinary collaboration and communication in all aspects of health care for humans, animals, and the environment. To overcome this disease of zoonoses, we developed a system of information zoonoses and Emerging Infectious Disease (SIZE). In this system of SIZE, we gather the disease information from social media. The disease information was collected from Twitter are Demam Berdarah Dengue (DBD), malaria disease, Antraks Disease, Canine Madness (Anjing Gila), Bird Flu (flu burung), and Ebola Disease. Twitter is a social media platform that has become a constant resource developing for data collectors. To perform this task to get the data of disease information, related tweets and Twitter user details the data collection using web scraping. Data Collection from Twitter was carried out by applying web scraping technology using python language. The scraping experiment from twitter in this study has succeeded in retrieving disease information from 2015-2020 using an advanced tool for Twitter scrapping called Twint using the python script. As the results lately have been increased number of tweets of diseases from canine madness (anjing gila) 34477, followed by Malaria Disease (28046) and Demam Berdarah Dengue (DBD) 11950 in 2020.

Keyword: web scraping, data collection, python, disease information, twint.

1. INTRODUCING

In recent years, the growth of social data on social networks or sites Web social networking has increased exponentially. This leads to the provision of data and information for scientists, academics, and researchers for many uses such as research, marketing, commercial, and academic. Social data on social networking sites contains many real actions that occur in everyday life [1].

One of the social media platforms that is widely used is Twitter. Twitter is a social media platform that has become a constant resource developing for data collectors. Twitter was first launched in July 2006 and when it is widely used by people all over the world. Twitter has a function as micro-blogging type of social media (small blog) with the number of characters in tweet (post in twitter) maximum 140 characters.

Crawling data on twitter is a process to retrieve or download data from the twitter server with the help of the good twitter Application Programming Integration (API) in the form of user data and tweet data. Crawling Twitter data using python is a one way of collecting text data. Nowadays not only text can get from twitter. Some other data can be known such as interaction time, number of likes, images, audio, and so on. In detail, the parts have been explained in the Twitter Application Programming Interface (API). The Twitter API or Twitter Application Programming Interface (API) is a programs or applications provided by twitter to make it easier for other developers in accessing information on the Twitter website.

Web scraping allows us to download data from many websites to our local system through the internet. It collects data from many internet sites utilizing HTTP or a web browser and analyzes it to meet our needs. Many researchers, companies utilized it to collect data and create a search engine. Beautiful soup and selenium are two python packages/modules that can aid in the process of web scraping. There are several libraries, such as Autoscraper, that can automate the web scraping process. All these libraries make use of different APIs for scraping data and storing it in a data frame on our local system.

There are many ways to crawling twitter data in python. Some are using tweepy, a library for interaction with the

JURNAL TEKNOINFO
Volume 16, Nomor 2, Juli 2022, Page 246-251
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

API of twitter. When used, tweepy requires developer accounts. when you don't have it then can't use tweepy as media to get twitter data. Overcoming this, one of them is by using twint. Twint is a tool used to get data from twitter. Twint developed using the python programming language and can be installed using pip or conda. Python language contains a number of useful packages that are suitable for geospatial analysis [2][3], suggested some implications deep learning using python [4], especially classification of micro-level countries [5], data analyses using jupyter notebook [6].

DBD is an infectious illness with a spectrum of clinical symptoms ranging from mild to severe. The symptoms of Dengue fever are followed by shock or dengue shock syndrome[7]. Malaria is an infectious illness caused by Plasmodium parasites (protozoa) that can be spread by the bite of a female Anopheles mosquito [8]. Anthrax is an acute infectious disease caused by the bacterium Bacillus anthracis and is a zoonotic disease. Anthrax disease mostly affects mammals and some bird species, especially herbivores. Livestock that are often contaminated are cows, buffaloes, goats, sheep, and pigs [9]. Rabies is a deadly and zoonotic virus illness that affects mammals, including humans. This illness infects the central nervous system, resulting in acute encephalomyelitis. Rabies virus (RV) and rabies-related viruses (RRVs) are members of the Rhabdoviridae family and the Lyssavirus genus [10]. Avian influenza (AI) is a deadly and highly infectious zoonotic illness that may infect all species of birds, people, pigs, horses, and dogs. The Drthomyxoviridae family's Avian Influenza Virus Type A (animals) has attacked humans and killed numerous people. Avian influenza is currently a major worldwide health concern, especially in Indonesia [11]. Ebola virus illness, often known as Ebola hemorrhagic fever (EHF), is an acute viral condition marked by fever and hemorrhage that has a high death rate in humans and nonhuman primates (primates) [12].

Environmental degradation caused by land conversion, trash (both domestic and industrial), and natural catastrophes are all variables that contribute to the establishment of disease susceptibility. The globe is facing a rising danger of novel infectious illnesses, often known as emerging infectious diseases (EID), 70 percent of which are zoonotic, or transmitted from animals to people [13]. Experts throughout the world suggest “ONE HEALTH” as a strategy for dealing with the threat of zoonoses. The One Health concept is a worldwide strategy to expand interdisciplinary collaboration and communication in all aspects of health care for humans, animals, and the environment. To overcome this disease of zoonoses, we developed a system information zoonoses and Emerging Infectious Disease (SIZE). In this system of SIZE, we gather the disease information from the social media. The disease information was collected from the twitter are Demam Berdarah Dengue (DBD), malaria disease, Antraks Disease, Canine Madness (Anjing Gila), Bird Flu (flu burung), and Ebola Disease.

The scraping experiment from twitter in this study has succeeded in retrieving disease information from 2015-2020 using an advanced tool for Twitter scrapping called Twint. Without needing to utilize the Twitter API, we can use this program to scrape any user’s followers, following, tweets, and so on. When the system are properly implemented with the disease information, it will help protect and save millions of lives in current and future generations.

2. RESEARCH METHODS

Crawling data is a stage in research that aims to collect or download data from a database. Data collection from research is downloaded from the Twitter server in the form of users and tweets along with their attributes [14]. Collecting data on rabies, bird flu and Ebola from social media twitter is done using a python script. Python scripts can be run through jupyter notebook or command prompt. Here are the steps in downloading data on Twitter social media:

2.1 Command prompt

To crawl data using python using the twint tools, the first thing to do is open a command prompt. Navigate the command prompt to the directory where we save the python script. After that, install twint by giving the command "pip install twint", press enter.

2.2 Jupyter notebook

Twint is a tool used to get data from twitter. Twint is developed using the python programming language and can be installed using pip or conda. Twint can be crape tweets based on many factors such as hastags, users, subjects, and so on. It can also extract information from tweets such as phone numbers and email addresses.
After the github installation is complete, then open the jupyter notebook. The jupyter notebook window will appear in the form of a tab in the browser that we use. In the jupyter notebook program, select Python 3, then enter the configuration script that will be used for crawling data from twitter.

JURNAL TEKNOINFO
Volume 16, Nomor 2, Juli 2022, Page 246-251
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

1. Importing libraries
   We will use Twint to scrape data from Twitter, therefor we will import Twint. Aside from that, we will need to import net asyncio, which will take care of any notebook or runtime problems. In addition, we shall solely use net syncio in this phase.
   ```
   import twint
   import nest_asyncio
   net_asyncio.apply()
   ```
2. Configuring Twint
   Before we can scrape data from Twitter with Twint, we must first configure the twint object and call it anytime it is needed.

2.3 Disease type configuration

For the disease type configuration (Table 1), we should configure such as:

1. Several types of diseases that are collected in the news on Twitter are Malaria, Dengue Fever and Anthrax. The keyword Malaria generates a large amount of data. This is because the use of the word applies globally.
2. Language
   A language filter is carried out where only tweets that are collected contain Indonesian.
3. Deadline
   Collected Tweets are also timed. The data taken is in the 2015-2020 period
4. Maximum Data Amount
   Data collection for each type of disease in one year is limited to a maximum number of 1,000,000 tweets.

Table 1. Script Command

Script	Details
import twint	
import nest_asyncio	
config = twint.Config()	
config.Search = “malaria”	disease type
config.Lang = “in”	Language type
config.Limit = 1000000	numbers of tweets maximum
config.Since = ‘2015-07-01 00:00:00’	data starts
config.Until = ‘2015-07-01 00:00:00’	data ends
config.Store_csv = True	
config.Output = “2015malaria.csv”	store data in csv file
nest_asyncio.apply()	
twint.run.Search(config)	

2.4 Data Analyses

Data analyses were processed using jupyter notebook and it is important to save the acquired data while getting new data in case the program stops unexpectedly. We can output data into different formats, including Pandas DataFrame, CSV, JSON, etc. To do so, we first need to define the output format, then list the filenames where the data is stored.
In my case I saved the data as a JSON file. Notice the "Hide_output" here. This command is very useful when we collect a lot of data and just want to save it in a file. If specified here, it will not display the scraped results in the program but save the data directly in the file.

3. RESULT AND DISCUSSIONS

Data collection was carried out to determine the number of interactions of EID disease that occurred on Twitter social media. The diseases for which data are sought are Dengue Berdarah Dengue (DBD), Malaria, Anthrax, Canine Madness,

JURNAL TEKNOINFO
Volume 16, Nomor 2, Juli 2022, Page 246-251
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Bird Flu and Ebola. The time range sought is in the period 2015 to 2020, so that later 18 disease interaction table data will be obtained.

3.1 Disease Type

1. Demam Berdarah Dengue (DBD)

Table 2. Number Tweet of DBD Disease

Year	Number data (Tweet)
2020	11950
2019	495
2018	6552
2017	10894
2016	34332
2015	39497
Total	103720

Data on Demam Berdarah Dengue (DBD) based on the number of tweets shows that in 2015 there were 39,497, in 2016 as many as 34,332. Tweet data in 2017 was 10,894, in 2018 there were 6,552, in 2019 as many as 495 and in 2020 as many as 11,950. There was a decrease in 2015-2019, but an increase again in 2020. The total data on DHF in the 2015-2020 period reached 103,720 tweets (Table 2).

2. Malaria Disease

Table 3. Number Tweet of Malaria Disease

Year	Number data (Tweet)
2020	28046
2019	15604
2018	13722
2017	14733
2016	22143
2015	31022
Total	125720

Malaria disease data based on the number of tweets shows a downward trend since 2015 as many as 31,022 tweets, in 2016 as many as 22,143 tweets, in 2017 as many as 14,733 tweets until 2018 to 13,722 tweets. Furthermore, there was an increase where in 2019 there were 15,604 tweets and in 2020 it became 28,406 tweets. The total number of Malaria tweets in Indonesia from 2015 to 2020 was 125,720 tweets (Table 3).

3. Antraks Disease

Anthrax disease data based on the number of tweets shows an increasing trend since 2015 as many as 2,395, in 2016 as many as 3,376, until 2017 to 7,075. Furthermore, there was a decrease where in 2018 there were 593 tweets. There was an increase again in 2019 by 888 and in 2020 with 1,458 tweets. The total number of Anthrax tweets in Indonesia from 2015 to 2020 is 16,145 tweets (Table 4).

Table 3. Number Tweet of Antraks Disease

Year	Number data (Tweet)
2020	1458
2019	888
2018	593
2017	7075
2016	3736
2015	2395
Total	16145

JURNAL TEKNOINFO
Volume 16, Nomor 2, Juli 2022, Page 246-251
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

4. Canine Madness (Anjing Gila), Bird Flu (flu burung), and Ebola Disease
   The total data for 2020 for rabid dogs is 36,826, with data in Indonesian as much as 34,477. The data for bird flu in 2020 is a total of 11,717 with data in Indonesian as much as 11,623. While the total data for Ebola in 2020 is 1,154,123 and those in Indonesian are 4,488. The downloaded data resume can be seen in the table below:

Table 4. Number Tweet of Canine Canine Madness, Bird Flu, and Ebola Disease

Year	Canine Madness	Bird Flu	Ebola
2020	34477	11623	4488
2019	15502	2496	2542
2018	6614	2125	4183
2017	5727	5259	3599
2016	6063	16982	4823
2015	11221	16787	38475
Total	79604	55272	58110

3.2. Data Analyses

The data analyses were processed the configuration of the type of disease. The output data into different formats, including Pandas DataFrame, CSV, JSON.

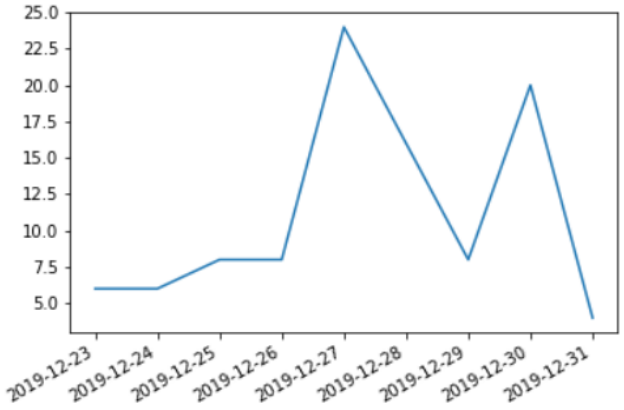

Figure 1. Number tweets by date of Bird Flu disease

4. CONCLUSION

This study demonstrates the tools of twint to perform this task to get the data of disease information, related tweets and twitter user details the data collection using web scraping. Crawling data is a stage in research that aims to collect or download data from a database. Data collection from research is downloaded from the Twitter server in the form of users and tweets along with their attributes. Data collection was carried out to determine the number of interactions of EID disease that occurred on Twitter social media. The diseases for which data are sought are Dengue Berdarah Dengue (DBD), Malaria, Anthrax, Canine Madness, Bird Flu and Ebola. The time range sought is in the period 2015 to 2020, so that later the disease interaction table data will be obtained. As the results, lately has been increased number of diseases from canine madness (anjing gila), followed by Malaria Disease and Demam Berdarah Dengue (DBD) in 2020. This study can be improved to ensure the other diseases and improve using the other model. The status of this research used by the number of tweets from the twitter and need to be validated in the surveillance by the ministry of health data.

JURNAL TEKNOINFO
Volume 16, Nomor 2, Juli 2022, Page 246-251
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

REFERENCES

[1] S. Kumar Singh, P. Verma, P. Kumar, and A. Abdul, "Journal of Critical Reviews Sentiment Analysis of Covid-19 Epidemic Using Machine Learning Algorithms on Twitter," vol. 7, no. 18, p. 2020, 2020.

[2] N. Nurda, R. Noguchi, and T. Ahamed, *Forest productivity and carbon stock analysis from vegetation phenological indices using satellite remote sensing in Indonesia*, vol. 4, no. 3. Springer Singapore, 2020.

[3] N. Nurda, R. Noguchi, and T. Ahamed, "Change detection and land suitability analysis for extension of potential forest areas in Indonesia using satellite remote sensing and GIS," *Forests*, vol. 11, no. 4, pp. 1–22, 2020, doi: 10.3390/F11040398.

[4] M. I. Habibie, T. Ahamed, and R. Noguchi, "Deep Learning Algorithms to determine Drought prone Areas Using Remote Sensing and GIS," *2020 IEEE Asia-Pacific Conf. Geosci. Electron. Remote Sens. Technol.*, doi: 10.1109/AGERS51788.2020.9452752.

[5] M. I. Habibie, R. Noguchi, S. Matsushita, and T. Ahamed, "Development of micro-level classifiers from land suitability analysis for drought-prone areas in Indonesia," *Remote Sens. Appl. Soc. Environ.*, vol. 20, no. June, p. 100421, 2020, doi: 10.1016/j.rsase.2020.100421.

[6] M. I. Habibie, R. Noguchi, M. Shusuke, and T. Ahamed, *Land suitability analysis for maize production in Indonesia using satellite remote sensing and GIS-based multicriteria decision support system*, vol. 5. Springer Netherlands, 2019.

[7] A. Candra, "Demam Berdarah Dengue : Epidemiologi , Patogenesis , dan Faktor Risiko Penularan Dengue Hemorrhagic Fever : Epidemiology , Pathogenesis , and Its Transmission Risk Factors," *Demam Berdarah Dengue Epidemiol. Patog. dan Fakt. Risiko Penularan*, vol. 2, no. 2, pp. 110–119, 2010.

[8] A. Ruliansyah and F. Y. Pradani, "Perilaku-Perilaku Sosial Penyebab Peningkatan Risiko Penularan Malaria di Pangandaran," *Bul. Penelit. Sist. Kesehat.*, vol. 23, no. 2, pp. 115–125, 2020, doi: 10.22435/hsr.v23i2.2797.

[9] C. Clarasinta and T. U. Soleha, "Penyakit Antraks : Ancaman untuk Petani dan Peternak," *Majority*, vol. 7, no. 1, pp. 158–164, 2017.

[10] M. Saepulloh and R. M. Abdul, "PEMETAAN GENETIK VIRUS RABIES PADA ANJING SEBAGAI DASAR PENETAPAN PENGENDALIAN PENYAKIT Genetic Mapping of Rabies Virus in Dogs as a Basis for Disease Control," *J. Kedokt. Hewan*, vol. 10, no. 1, pp. 43–48, 2010.

[11] F. Elytha, "Sekilas Tentang Avian Influenza (Ai)," *J. Kesehat. Masy.*, vol. 6, no. 1, pp. 47–50, 2006.

[12] H. E. Yanti and Aryati, "Penyakit Virus Ebola (Ebola Virus Disease)," *Indones. J. Clin. Pathol. Med. Lab.*, vol. 21, no. 2, pp. 195–201, 2015, [Online]. Available: file:///C:/Users/A4/AppData/Local/Temp/351-178-1-SM.pdf.

[13] Kementerian Koordinator Bidang Pembangunan Manusia dan kebudayaan, "Implementasi one health di indonesia," p. 2, 2016.

[14] J. Eka Sembodo, E. Budi Setiawan, and Z. Abdurahman Baizal, "Data Crawling Otomatis pada Twitter," no. September, pp. 11–16, 2016, doi: 10.21108/indosc.2016.111.