E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1211 - 1219

WEB SCRAPING WITH HTML DOM METHOD FOR WEBSITE NEWS API CREATION

Maulana Irfan Firdian[1)], Eko Darwiyanto[2)], and Monterico Adrian[3)]
[1, 2, 3)]Faculty of Informatics, Telkom University, Bandung
Jl. Telekomunikasi No. 1, Sukapura, Kec. Dayeuhkolot, Kabupaten Bandung, Jawa Barat 40257
e-mail: maulanairfanf@students.telkomuniversity.ac.id [1)], ekodarwiyanto@telkomuniversity.ac.id[2)], monterico@telkomuniversity.ac.id[3)]

ABSTRAK

Informasi merupakan salah satu hal yang penting pada era sekarang, salah satu informasi yang selalu ada setiap harinya adalah berita. Banyaknya berita-berita yang muncul setiap harinya menjadi sebuah masalah baru ketika website berita tidak menyediakan layanan API (Application Programming Interface) untuk mendapatkan berita-berita tersebut. Hal tersebut menjadi kendala bagi peneliti yang akan melakukan analisis topik berita. Cara salin dan tempel kurang efektif digunakan untuk mendapatkan berita setiap harinya pada website berita dikarenakan membutuhkan waktu yang cukup lama. Pada penelitian ini dilakukan web scraping dengan metode HTML (Hypertext Markup Language) DOM (Document Object Model) untuk mengambil data dari situs berita. Hasil web scraping berupa dataset yang kemudian dimasukkan kedalam database dan dijadikan sebuah API. API yang telah dibuat dilakukan pengujian menggunakan black box testing dan dilakukan pengujian kesesuian datanya, antara data yang diperoleh pada saat melakukan scraping dan data yang ada pada website berita pada saat dilakukan pengujian. Hasil pengujian menggunakan black box testing menunjukkan bahwa filter-filter pada API yang dibuat berjalan sesuai fungsinya dan mendapatkan persentase kesesuaian data yang tinggi. Pada website berita Tribunnews.com memiliki tingkat kesesuaian sebesar 99,2%, Detik.com sebesar 97,9% dan Liputan6.com sebesar 98,6.%.

***Kata Kunci**: API, Berita, HTML DOM, Website, Web scraping*

ABSTRACT

Information is one of the important things in this era, one of the information that always exists every day is news. The amount of news that appears every day becomes a new problem when news websites do not provide API (Application Programming Interface) services to get the news. This is an obstacle for researchers who will analyze news topics. The copy and paste method is less effective in getting news every day on news websites because it takes a long time. In this research, web scraping is done with the HTML (Hypertext Markup Language) DOM (Document Object Model) method to retrieve data from news sites. The results of web scraping are in the form of datasets which are then entered into the database and made into an API. The API that has been created is tested using black box testing and testing the suitability of the data, between the data obtained when scraping and the data on the news website at the time of testing. The results of testing using black box testing show that the filters on the API created run according to their functions and get a high percentage of data conformity. The Tribunnews.com news website has a conformity rate of 99.2%, Detik.com of 97.9% and Li-putan6.com of 98.6%.

***Keywords**: API, HTML DOM, News, Website, Web Scraping*

I. INTRODUCTION

INFORMATION is one of the most important things in the current era, thus encouraging humans to develop technology that can process data quickly and efficiently. People can easily get information and news that is read to then become a reference in life. This has become one of the causes of the increasing number of news websites [1]. Information about various kinds of news is one of the many information that can be accessed through sites such as Liputan6.com, Tribunnews.com, Detik.com and so on. Quoted from Alexa Traffic Rank (ATR) there are 8 news websites out of 10 websites with the highest access in Indonesia.

The amount of new news that appears every day becomes a new problem when news websites do not provide API services to download these news. The copy and paste method cannot be used to get news from news websites every day because it will take a very long time [2]. Web scraping technique can be a solution to the problem because this technique can retrieve data from a website quickly. Web scraping is the process of retrieving semi-structured data from the internet, which is generally in the form of web pages in markup languages such as HTML (Hypertext Markup Language) or XHTML (Extensible Hypertext Markup Language), and analyzing these documents to retrieve certain data from the page to be used for other purposes [3]. The application of web scraping has been carried out by many previous studies including research on the implementation of web scraping in retrieving and collecting data from google scholar scientific articles [4], the implementation of Web Scraping in the collection of criminal

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1211 - 1219

news during the Covid-19 Pandemic to determine the trend in the number of criminal news [5], the implementation of web scraping used to retrieve information on e-commerce sites, e-marketplace in the form of product description content in several e-commerce [6] and e-marketplace sites and the use of web scraping to compare the prices of components in making computers from several computer shop websites [7].

Based on previous research, it is known that web scraping can facilitate the data retrieval process, such as in the research Web Scraping with HTML DOM Method for Data Collection of Scientific Articles from Google Scholar [8] and An Approach of Web Scraping on News Website based on Regular Expression. However, not applying an automatic scheduler in web scraping is also a limitation in these studies because it does not obtain data every time there is a change. In addition, the data from the web scraping results in the study was not published so that other researchers who need the data must do web scraping again because they did not get access to data from the study. The absence of API services on news websites is a problem for researchers when they want to get data from the website so that it is necessary to have an API to make it easier to get the desired data. This is supported by research entitled "Sentiment Analysis of the News Media on Artificial Intelligence Does Not Support Claims of Negative Bias Against Artificial Intelligence" that the data used in the study came from the New York Times news website API [9] and "Detection Of Online Fake News: A Survey" which uses data from the Facebook graph API to perform classification [10].

The results of web scraping on the news website will be used to create an API and test the API that has been created to find out whether the API is good enough to use. The HTML DOM method is used in this research because based on previous research the method has an average scraping time that is faster and uses less internet compared to the other two methods, namely Regular Expression and Xpath [11].

II. RESEARCH METHOD

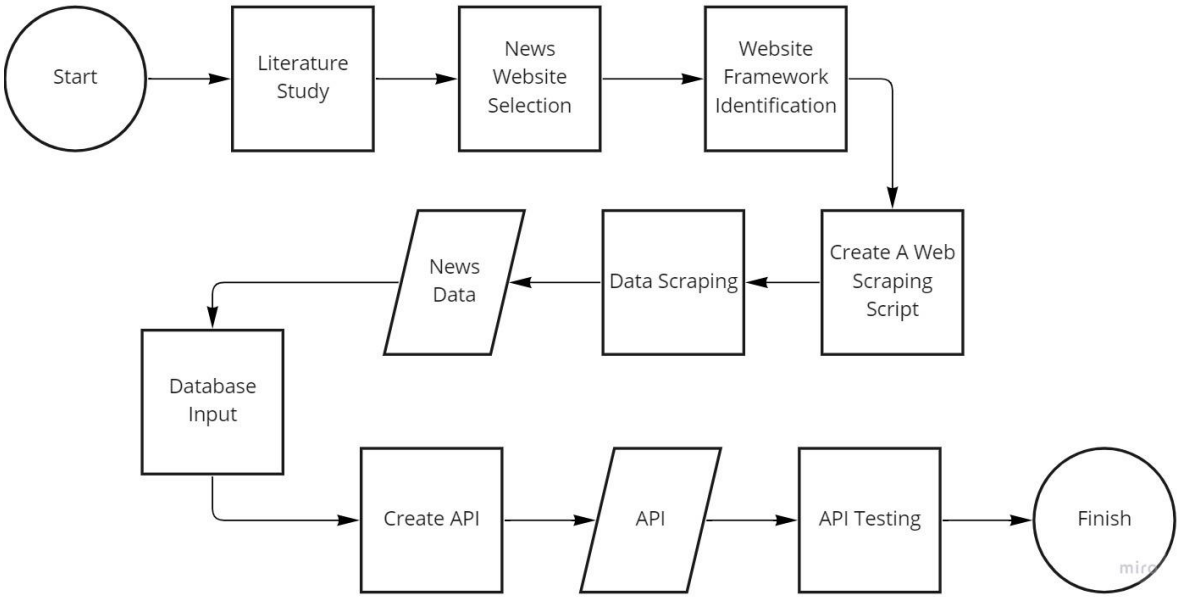


Fig 1. Stages of Work

The flow of the research stages is carried out in several stages, namely:

1) *Literatur Study*

Studies carried out to fulfill the theory that has been obtained include learning web scraping programming techniques, HTML DOM, databases, websites and FastAPI Framework through the internet, articles, books and journals. The results obtained are used as a reference for this research.

2) *News Website Selection*

The selection of news websites is based on the journal that is the reference for the author to conduct research. The websites include Liputan6.com, Detik.com and Tribunnews.com. The reason for choosing the three news websites is because the three news websites have a high level of access in Indonesia [2].

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1211 - 1219

3) Website Framework Identification

This stage is done as a reference in creating a script for scraping. At this stage the author will go to the website page that will be scraped, and see the framework on the website using the inspect element. The HTML structure on each news website is quite different between one news website and another, so it is necessary to identify the website framework of each news website that will be scraped [4]. To get the desired news detail link, the author needs to find the location of the <a> tag on the main news page that goes to the news detail page. After successfully obtaining the news detail link, the author re-identifies the framework on the news detail page of each news website to get the required data in the form of title, author, publication date and news content. After completing the identification of the framework on the news detail page, the author creates a customized script on each news detail page framework to be able to retrieve predetermined data.

4) Creating Scraping Web Scripts

At this stage the author creates a script for scraping using the python programming language and the help of the beautifulsoup library which is used to pull data out of HTML and XML files [12]. The scraping script for each news website is different because from the identification of the website framework, each website has a different structure. The scraping script will be adjusted to the framework on each news website page. After successfully scraping, text processing is carried out which aims to create a format for the date of publication of the news, eliminate text containing advertisements and eliminate blank spaces [5].

5) Scraping Data

At this stage, the author will use the scraping script that has been made previously to get news data from a predetermined website. The result of scraping is a dataset which will then be processed for research purposes. Scraping this data is done automatically every 2 hours every day using the help of the WaysSript.com website, the author needs to upload the scraping code file that has been made, fill in the commands to run the code and set the time to do the scraping.

6) *Scraping Result*

At this stage, an analysis is carried out on the data that has been obtained through the previous stage, namely scraping data. The data obtained has clean results, there are no sentences containing advertisements such as those on the website, there are no long blank spaces, and there are no unnecessary elements such as HTML tags. The results of date formatting on each website already have the same format, namely (DD-MM-YYYY), the purpose of date formatting is to make it easier to filter the date of publication of news on the API.

7) Database Input

The next stage after creating and filling data into the database is API creation. API creation is done using the FastAPI framework to provide access to data that has been collected so that it can be used by the public for their respective purposes. The API created has several filters, namely searching by news title, news publication date, news author, news website origin, news category and a combination of the filters mentioned earlier. So that the API that has been created can be accessed by the public, the author is hosting the API that has been created on the Heroku server.

8) API Creation

The next stage after creating and filling data into the database is API creation. API creation is done using the FastAPI framework to provide access to data that has been collected so that it can be used by the public for their respective purposes. The API created has several filters, namely searching by news title, news publication date, news author, news website origin, news category and a combination of the filters mentioned earlier. So that the API that has been created can be accessed by the public, the author is hosting the API that has been created on the Heroku server.

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1211 - 1219

9) API Testing

At this stage, 2 tests are carried out, the first is black box testing to check whether the features in the API are running properly and try to find unexpected errors through Swagger UI FastAPI [13] [14]. Then testing is carried out to see the suitability of the headlines in the API and the headlines on the news website when testing.

III. RESULT & DISCUSSION

Scraping Results

The scraping results are in the form of a dateset that has been text processed and some adjustments so that the data from the three news websites has the same date format and produces news that is clean from blank spaces and unnecessary advertisement text [5]. The following are the results of web scraping on 3 news websites conducted on June 3, 2022. The scraping results from each website can be seen in Figure 2, Figure 3 and Figure 4.

title	date	author	link	ategor	website	content
Jadwal Ler	03-06-202	Gilar Ram	https://www	biasa	liputan6	Liputan6.com,
Aktor Kris	03-06-202	Ady Anugr	https://www	biasa	liputan6	Liputan6.com,
The Amba	03-06-202	Tanti Yulia	https://www	biasa	liputan6	Liputan6.com,
Besok, Jok	03-06-202	Lizsa Egeh	https://www	biasa	liputan6	Liputan6.com,
Tim Penye	03-06-202	Wayan Dia	https://www	biasa	liputan6	Liputan6.com,
FOTO: KPK	03-06-202	Johan Fatz	https://www	biasa	liputan6	Haryadi Suyuti
VIDEO: Le	03-06-202	Gilang Faj	https://www	biasa	liputan6	Liputan6.com,
Infografis	03-06-202	Shinta NM	https://www	biasa	liputan6	Liputan6.com,
Bola Ganji	29-05-202	Harley Ikh	https://www	biasa	liputan6	Liputan6.com,

Fig 2. Some scraping data from the Liputan6.com website

title	date	author	link	category	website	content
BREAKING	03-06-202	Taufik Ism	https://wv	biasa	tribunnew	Laporan Wart
Kronologi	03-06-202	Adi Suhen	https://wv	biasa	tribunnew	TRIBUNNEWS
Pria yang	03-06-202	Author tid	https://wv	biasa	tribunnew	Laporan Wart
KIB Bakal	03-06-202	Rizki Sandi	https://wv	biasa	tribunnew	Laporan Repo
Soal Inova	03-06-202	Muhamma	https://wv	biasa	tribunnew	TRIBUNNEWS
Sampai Pe	03-06-202	Author tid	https://wv	biasa	tribunnew	Sampai Pensiu
Kepala BP	03-06-202	Hasanudin	https://wv	biasa	tribunnew	TRIBUNNEWS
Satu Penu	03-06-202	Author tid	https://wv	biasa	tribunnew	Laporan Wart
Menpan-R	03-06-202	Author tid	https://wv	biasa	tribunnew	TRIBUNNEWS

Fig 3. Some scraping data from the Tribunnews.com website

title	date	author	link	category	vebsit	content
Ahmad Sa	03-06-202	Tiara Aliya	https://ne	popular	detik	Ketua Pelaksan
Keji! Pria d	03-06-202	Mei Ameli	https://ne	popular	detik	Polsek Cengkar
Potret Rid	03-06-202	Tim detikc	https://ne	popular	detik	Jakarta - Ridwa
Jokowi Bal	03-06-202	Tiara Aliya	https://ne	popular	detik	Presiden Joko \
Pesan Sop	03-06-202	Pingkan A	https://ho	popular	detik	Eva Celia dan D
Hasil FP2 M	03-06-202	Okdwitya	https://sp	popular	detik	Aleix Espargarc
Kronologi	03-06-202	Rosmha W	https://ww	popular	detik	Putra Sulung G
Johnny De	03-06-202	Daniel Nga	https://wc	popular	detik	Dukungan men
Kisah Bule	03-06-202	Siti Fatima	https://tra	popular	detik	Ada pemandar

Fig 4. Some scraping data from the Detik.com website

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1211 - 1219

Web Scraping Test Results

In this test, web scraping is done 10 times from each news website with a time lag of 30 minutes. In this study, a comparison of web scraping testing using another method, CSS Selector, was also conducted. The purpose of this test is to see the scraping time, the amount of data obtained and memory usage when scraping from the two methods used. The results of the comparison of the two methods can be seen in Table I, Table II, and Table III.

TABLE I
COMPARISON OF SCRAPING RESULTS ON THE TRIBUNNEWS.COM WEBSITE

Experiment	HTML DOM			CSS Selector		
	Count Data	Time (s)	Memory (bytes)	Count Data	Time (s)	Memory (bytes)
1	60	8,52	17.854.753	65	9,77	17.968.453
2	61	9,5	17.818.883	64	9,98	18.048.425
3	62	7,76	17.964.184	65	9,88	18.109.164
4	51	8,49	17.747.012	55	10,77	17.940.997
5	52	7,64	17.850.699	56	9,90	17.993.936
6	56	7,70	17.963.407	59	9,77	17.998.246
7	57	8,24	17.566.480	62	10,01	17.730.205
8	60	8,28	17.984.173	65	10,07	17.806.396
9	58	8,68	17.646.572	63	9,99	17.769.754
10	58	9,48	17.553.590	61	10,79	17.768.663
Average	57,5	8,43	17.794.975,3	61,5	10,09	17.913.423,9

Table I shows the comparison of scraping results on the Tribunnews.com website, where the scraping results with the HTML DOM method obtained an average amount of data of 57.5, scraping time of 8.43 seconds and memory usage of 17,794,975.3 bytes. Meanwhile, the CSS Selector method obtained an average amount of data of 61.5, scraping time of 10.09 and memory usage of 17,913,423.9 bytes.

TABLE II
COMPARISON OF SCRAPING RESULTS ON THE LIPUTAN6.COM WEBSITE

Experiment	HTML DOM			CSS Selector		
	Count Data	Time (s)	Memory (bytes)	Count Data	Time (s)	Memory (bytes)
1	82	61,07	23.791.143	83	65	25.309.752
2	82	61,80	23.989.896	82	64,43	23.971.855
3	80	59,82	24.897.734	80	66,50	23.610.221
4	78	58,69	23.501.145	78	63,27	23.564.908
5	74	63,24	24.500.103	75	64,94	23.439.782
6	70	54,91	23.870.166	70	58,66	24.736.864
7	66	53,17	25.120.500	66	55,19	25.049.178
8	64	53,67	24.964.602	65	57,52	24.892.059
9	68	55,93	24.944.788	67	57,15	24.634.559
10	71	55,48	24.203.144	71	56,29	24.801.181
Average	73,5	57,78	24.378.322,1	73,7	60,9	24.401.035,9

Table II shows the comparison of scraping results on the Liputan6.com website, where the results of scraping with the HTML DOM method obtained an average amount of data of 73.5, scraping time of 57.78 and memory usage of 24,378,322.1 bytes. Meanwhile, the CSS Selector method obtained an average amount of data of 73.7, scraping time of 60.9 seconds and memory usage of 24,401,035.9 bytes

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1211 - 1219

TABLE III
COMPARISON OF SCRAPING RESULTS ON THE DETIK.COM WEBSITE

Experiment	HTML DOM			CSS Selector		
	Count Data	Time (s)	Memory (bytes)	Count Data	Time (s)	Memory (bytes)
1	67	159,60	20.737.888	67	158,97	20.809.277
2	71	149,45	20.631.085	70	151,22	20.704.027
3	68	141,32	20.512.194	69	147,03	20.550.642
4	68	153,22	20.499.850	68	154,73	20.579.439
5	71	163,60	20.367.359	71	163,38	20.447.549
6	69	153,39	20.444.721	71	152,16	20.430.366
7	68	158,28	19.626.534	69	146,67	19.862.024
8	69	159,01	20.545.795	70	159,04	20.621.024
9	69	150,85	20.584.461	69	148,51	20.260.957
10	69	147,27	20.622.541	69	159,62	20.675.745
Average	68,9	153,60	20.457.242,8	69,3	154,13	20.494.105

Table 3 shows the comparison of scraping results on the Detik.com website, where the results of scraping with the HTML DOM method obtained an average amount of data of 68.9, scraping time of 153.60 seconds and memory usage of 20,457,242.8 bytes. Meanwhile, the CSS Selector method obtained an average amount of data of 69.3, scraping time of 154.13 seconds and memory usage of 20,494,105 bytes.

Based on the comparison of scraping results using these 2 methods, it can be concluded that the HTML DOM method has a faster average scraping time, smaller memory usage, and less data compared to the CSS Selector method. Testing web scraping on the Tribunnews.com website produces the fastest scraping time because the framework on the Tribunnews.com website is not too complicated compared to the other 2 websites. Furthermore, web scraping testing on the Liputan6.com website produces the most amount of data compared to the other 2 websites due to the more news available on the website. As for memory usage, it is directly proportional to the amount of data obtained on each news website.

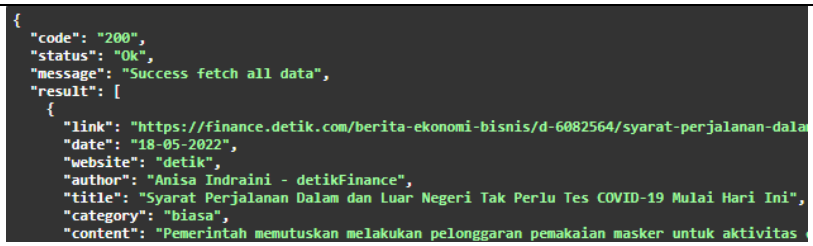
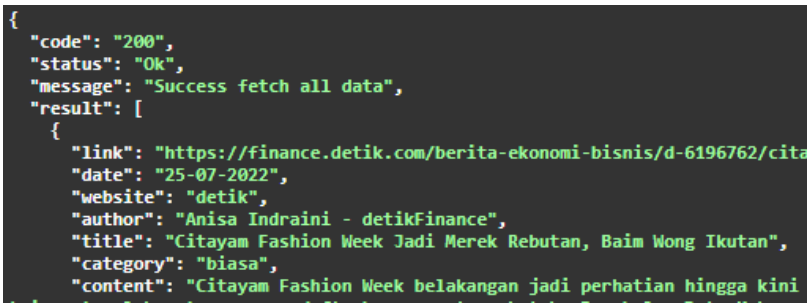
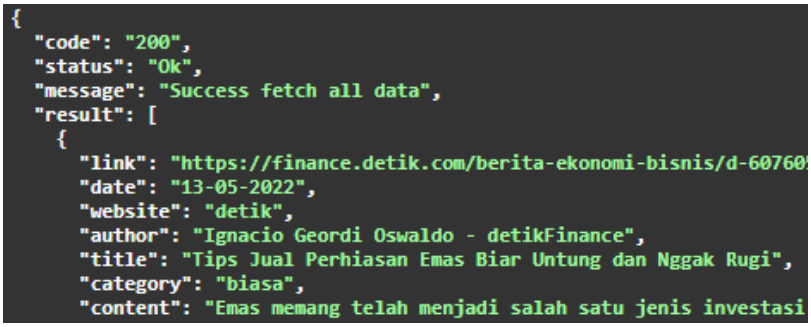
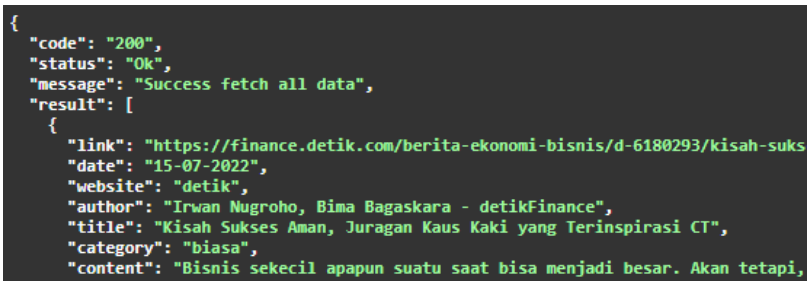
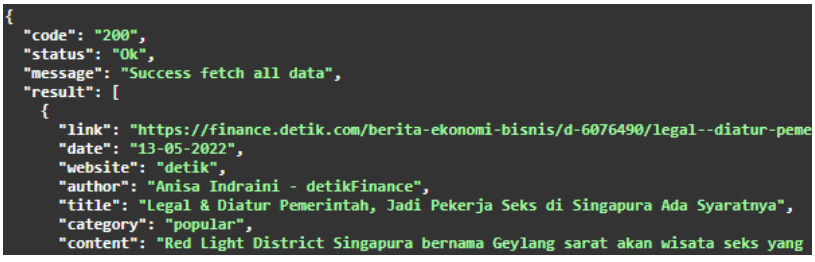
Functional API Test Results

The API that has been created is tested, this test is carried out using Black Box Testing. This is done by using all the filters available on the API. In addition, this test is carried out to check whether there are code errors that have been made by the author. The results of the filters in the API have run quite well, getting data that matches the parameters given, and API performance is also fast enough by getting an average time of 7.5 seconds to retrieve 1000 data. The results of functional API testing can be seen in Table IV.

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1211 - 1219

TABLE IV
API FUNCTIONAL TESTING RESULTS

Test Explanation	Expected Results	Testing Results	Description
Search for news based on news headlines (covid-19)	API displays news data that contains the given news title	 { "code": "200", "status": "Ok", "message": "Success fetch all data", "result": [{ "link": "https://finance.detik.com/berita-ekonomi-bisnis/d-6082564/syarat-perjalanan-dala "date": "18-05-2022", "website": "detik", "author": "Anisa Indraini - detikFinance", "title": "Syarat Perjalanan Dalam dan Luar Negeri Tak Perlu Tes COVID-19 Mulai Hari Ini", "category": "biasa", "content": "Pemerintah memutuskan melakukan pelonggaran pemakaian masker untuk aktivitas	Successful
Search for news based on the date the news was published (25-07-2022)	API displays news data according to the given news date	 { "code": "200", "status": "Ok", "message": "Success fetch all data", "result": [{ "link": "https://finance.detik.com/berita-ekonomi-bisnis/d-6196762/cita "date": "25-07-2022", "website": "detik", "author": "Anisa Indraini - detikFinance", "title": "Citayam Fashion Week Jadi Merek Rebutan, Baim Wong Ikutan", "category": "biasa", "content": "Citayam Fashion Week belakangan jadi perhatian hingga kini	Successful
Searching for news based on the origin of the news website (detik)	API displays news data according to the given website	 { "code": "200", "status": "Ok", "message": "Success fetch all data", "result": [{ "link": "https://finance.detik.com/berita-ekonomi-bisnis/d-60760 "date": "13-05-2022", "website": "detik", "author": "Ignacio Geordi Oswaldo - detikFinance", "title": "Tips Jual Perhiasan Emas Biar Untung dan Nggak Rugi", "category": "biasa", "content": "Emas memang telah menjadi salah satu jenis investasi	Successful
Search for news based on the name of the news writer (bagas)	API displays news data containing the given author name	 { "code": "200", "status": "Ok", "message": "Success fetch all data", "result": [{ "link": "https://finance.detik.com/berita-ekonomi-bisnis/d-6180293/kisah-suks "date": "15-07-2022", "website": "detik", "author": "Irwan Nugroho, Bima Bagaskara - detikFinance", "title": "Kisah Sukses Aman, Juragan Kaus Kaki yang Terinspirasi CT", "category": "biasa", "content": "Bisnis sekecil apapun suatu saat bisa menjadi besar. Akan tetapi,	Successful
Search for news based on news categories (popular)	API displays news data according to the given category	 { "code": "200", "status": "Ok", "message": "Success fetch all data", "result": [{ "link": "https://finance.detik.com/berita-ekonomi-bisnis/d-6076490/legal--diatur-peme "date": "13-05-2022", "website": "detik", "author": "Anisa Indraini - detikFinance", "title": "Legal & Diatur Pemerintah, Jadi Pekerja Seks di Singapura Ada Syaratnya", "category": "popular", "content": "Red Light District Singapura bernama Geylang sarat akan wisata seks yang	Successful
Perform a search based on a combination of several parameters (citayam fashion week, 19-07-2022, detik, popular, linda)	The API displays news data according to several parameters given	 { "code": "200", "status": "Ok", "message": "Success fetch all data", "result": [{ "link": "https://news.detik.com/berita/d-6187980/catwalk-di-spot-citayam-fashion-week- "date": "19-07-2022", "website": "detik", "author": "Marlinda Oktavia Erwanti - detikNews", "title": "Catwalk di Spot Citayam Fashion Week, Anies: Kami Tak Sekeren Mereka", "category": "popular", "content": "Gubernur DKI Jakarta Anies Baswedan mengajak Wakil Presiden Bank Investasi	Successful

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1211 - 1219

API Data Conformity Test

Testing is done by taking 30.000 data from the API that has been created 10.000 each from each news website, then each news title that comes from the API will be compared with each news title on the website when testing. The results of the test are differences such as uppercase and lowercase letters at the beginning of words, changes (subtraction / addition) of several words, and changes in (addition / subtraction) punctuation. The following are some comparisons between the headlines in the API and the headlines on the news website during validation and the percentage of matches between the headlines in the API and the headlines on the news website during validation, which can be seen in Table V, Table VI and Table VII.

TABLE IV
COMPARISON OF TITLES ON THE API AND TRIBUNNEWS.COM WEBSITE

No	API	Website	Explanation
1	Menko PMK: Tidak Ada Peningkatan Signifikan Kasus Covid-19 Usai Mudik 2022	Menko PMK: Tidak Ada Peningkatan Signifikan Kasus Covid 19 Usai Mudik 2022	Reducing punctuation on the word “Covid-19” to “Covid 19”
2	Tarif Listrik Pelanggan 3.000 VA Bakal Naik, PLN Bilang Begini	Jokowi Setuju Tarif Listrik 3.000 VA Naik, PLN Bilang Begini	Change a few words
3	Kisah Mbah Amir Warga Kudus Terpaksa Tinggal di Kudus Karena Tidak Dirawat Keluarga	Kisah Mbah Amir Warga Kudus Terpaksa Tinggal di Gubuk Dekat Toilet Karena Tidak Dirawat Keluarga	Change a few words
4	5 Alasan Liverpool Ngebet Boyong Jarrod Bowen dari West Ham, Salah Satunya Pengganti Sepadan Salah	5 Alasan Liverpool Ngebet Boyong Jarrod Bowen dari West Ham, Diprediksi Pengganti Sepadan Mo Salah	Change a few words
5	Hasil Final Thomas Cup 2022 Hari Ini - Indonesia Tak Beruntung, India Segel Gelar Juara Perdana	Hasil Final Thomas Cup 2022 Hari Ini - Indonesia Tak Beruntun, India Segel Gelar Juara Perdana	Change the word "Beruntung" to “Beruntun”

TABLE VI
COMPARISON OF TITLES ON THE API AND ON THE TRIBUNNEWS.COM WEBSITE

No	API	Website	Explanation
1	Komedo dan Jerawat Lenyap dalam 30 Menit, Begini Treatmentnya	Komedo dan Jerawat Lenyap dalam 30 Menit, Begini Treatment Salon Kecantikan di Surabaya	Change a few words
2	Polisi Tangkap Pembunuh Wanita di Kediri, Begini Ceritanya	Polisi Tangkap Pria Pembunuh Wanita Teman Kencan di Kediri, Begini Ceritanya	Change a few words
3	Update Pencarian Anak Ridwan Kamil di Sungai Aare, Tim SAR Kesulitan Karena Ini	Update Pencarian Anak Ridwan Kamil di Sungai Aare, Tim SAR Kesulitan karena Ini	Change of capital letters from “Karena” to “karena”
4	Sri Mulyani Pastikan APBN 2023 Masih Defisit di Kisaran Minus 2,61 Persen PDB	APBN 2023 Masih Defisit di Kisaran Minus 2,61 Persen PDB	Change a few words
5	Sri Mulyani Tambah Subsidi Energi Rp 350 Triliun untuk Cegah Kenaikan Harga Pertalite	Sri Mulyani Tambah Subsidi dan Kompensasi Energi Rp 350 Triliun untuk Cegah Kenaikan Harga Pertalite	Change a few words

TABLE VII
COMPARISON OF TITLES ON THE API AND ON THE DETIK.COM WEBSITE

No	API	Website	Explanation
1	Hadir di Jaksel, Southgate Residence Beri 5 Keuntungan bagi Penghuni	Hadir di Jaksel, Southgate Residence Beri 4 Keuntungan bagi Penghuni	Change number from number "5" to "4"
2	Pencinta Kopi! 5 Kopi Legendaris di Bandung Ini Layak Dicoba	Pencinta Kopi! 5 Tempat Ngopi Legendaris di Bandung Ini Layak Dicoba	Change a few words
3	Congo yang Masuk Daftar Negara Termiskin di Dunia, Begini Suasana Pasar Tradisionalnya	Kongo yang Masuk Daftar Negara Termiskin di Dunia, Begini Suasana Pasar Tradisionalnya	Change of capital letters from “Congo” to “Kongo"
4	Mata Sama Otak Lagi Nggak Fokus? Asah Dengan Teka-Teki Ini	Mata Sama Otak Lagi Nggak Fokus? Asah dengan Teka-teki Ini	Change of capital letters from “Dengan” to “dengan”
5	Dinkes DKI Luruskan Pernyataan Wagub Soal 14 Kasus Konfirmasi Hepatitis Akut	Dinkes DKI Jelaskan Pernyataan Wagub Soal 14 Kasus Konfirmasi Hepatitis Akut	Change a few words

After testing between the news title data on the API and the news website, a fairly high percentage of matches was obtained. The results of the percentage of data matches from the API and the website can be seen in Table VIII.

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1211 - 1219

TABLE VII
PERCENTAGE OF MATCH DATA FROM API AND NEWS WEBSITE

Website	Data	Change	Percentage
Liputan6.com	10000	145	98.6%
Tribunnews.com	10000	82	99,2%
Detik.com	10000	214	97.9%

IV. CONCLUSION

Based on the results and tests that have been carried out, the following conclusions are obtained that web scraping with the HTML DOM method on the three news websites can be done well using the python programming language. After scraping, data cleaning of scraping results is also carried out for research needs and API creation.

The news website API that has been created is functioning properly, this is directly proportional to the results of black box testing, which has the result that the filters available on the API display news data according to the keyword given. The level of title validation between the news provided by the API and the news provided on the website is quite high, namely 98.6% on the Liputan6.com website, 99.2% on the Tribunnews.com website and 97.9% on the Detik.com website. The difference between the news on the API and the news on the news website is due to news updates from the news website.

REFERENCES

[1] M. A. Adli and L. Firgia, "Rancang Bangun Web Scraping Pada Media Online Berita Nasional," *Jurnal ENTER*, vol. 1, pp. 118-128, 2018.

[2] A. Maududie, W. E. Y. Retnani and M. A. Rohim, "An Approach of Web Scraping on News Website based on Regular Expression," in *The 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, Makassar, 2018.

[3] M. Turland, Php|architect's Guide to Web Scraping, Marco Tabini & Associates, Inc., 2010.

[4] A. Josi, L. A. Abdillah and Suryayusra, "Penerapan Teknik Web Scraping Pada Mesin Pencari Artikel Ilmiah," *Jurnal Sistem Informasi (SISFO)*, vol. 5, pp. 1-6, 2014.

[5] S. Satriajati, S. B. Panutun and S. Pramana, "Implementasi Web Scraping dalam Pengumpulan Berita Kriminal pada Masa Pandemi Covid-19," in *Seminar Nasional Official Statistics*, Jakarta, 2020.

[6] D. F. Setiawan, Tristiyanto and A. Hijriani, "APLIKASI WEB SCRAPING DESKRIPSI PRODUK," *Jurnal TEKNOINFO*, vol. 14, no. 1, pp. 41-47, 2020.

[7] L. R. Julian and F. Natalia, "THE USE OF WEB SCRAPING IN COMPUITER PARTS AND ASSEMBLY PRICE COMPARISON," in *3rd International Conference on New Media (CONMEDIA)*, Tangerang, 2015.

[8] A. Rahmatulloh and R. Gunawan, "Web Scraping with HTML DOM Method for Data Collection of Scientific Articles from Google Scholar," *Indonesian Jurnal of Information Systems (IJIS)*, vol. 2, no. 2, pp. 95-104, 2020.

[9] C. M. Colin Garvey, "Sentiment Analysis of the News Media on Artificial Intelligence Does Not Support Claims of Negative Bias Againts Artificial Intelligence," *A Journal of Integrative Biology*, vol. 23, no. 0, pp. 1-14, 2019.

[10] S. Gaonkar, S. Itagi, R. Chalippatt, A. Gaonkar, S. Aswale and P. Shetgaonkar, "Detection Of Online Fake News : A Survey," in *International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*, Vellore, 2019.

[11] R. Gunawan, A. Rahmatulloh, I. Darmawan and F. Firdaus, "Comparison of Web Scraping Techniques: Regular Expression, HTML DOM and Xpath," in *International Conference on Industrial Enterprise and System Engineering*, Yogyakarta, 2018.

[12] L. Richardson, "Beautiful Soup," [Online]. Available: https://beautiful-soup-4.readthedocs.io/en/latest/. [Accessed 2022 07 26].

[13] D. D. Ayani, H. S. Pratiwi and H. Muhardi, "Implementasi Web Scraping untuk Pengambilan Data pada Situs Marketplace," *Jurnal Sistam dan Teknologi Informasi*, vol. 7, no. 4, pp. 257-262, 2019.

[14] M. Iskhak and S. Rizkika, "Implementasi Metode Pengujian Equivalence Partitioning pada pengembangan RESTful API Sistem Informasi Klinik Pratama UPN "Veteran" Yogyakarta," in *SEMNASIF*, Yogyakarta, 2021.

[15] S. N. Yanti and E. Rihyanti, "Penerapan Rest API untuk Sistem Informasi Film Secara Daring," *Jurnal Informatika Universitas Pamulang*, vol. 6, no. 1, pp. 195-201, 2021.

[16] Hasanuddin, H. Asgar and B. Hartono, "Rancang Bangun REST API Aplikasi Weshare Sebagai Upaya Mempermudah Pelayanan Donasi Kemanusiaan," *Jurnal Informatika Teknologi dan Sains*, vol. 4, no. 1, pp. 8-14, 2022.