

# ANALISIS WEB SCRAPING UNTUK DATA BENCANA ALAM DENGAN MENGGUNAKAN TEKNIK BREADTH-FIRST SEARCH TERHADAP 3 MEDIA ONLINE

*Izatul Putri Sonya*<sup>1</sup>  
*Dr. Prihandoko, SKom*<sup>2</sup>

<sup>1,2</sup>Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Gunadarma  
<sup>2</sup>prihandoko@staff.gunadarma.ac.id

## Abstrak

Teknologi yang ada saat ini memungkinkan seseorang untuk bekerja dengan mudah dan cepat. Salah satunya dalam hal memperoleh informasi dari web. Proses pengambilan informasi dari situs-situs web disebut dengan *web scraping*. Pada penelitian ini dilakukan analisis *web scraping* terkait bencana alam dari 3 situs media *online*, yaitu Detikcom, Liputan6, dan VivaNews. Fokus *web scraping* lebih kepada data yang tidak terstruktur pada web, menjadi sebuah data yang dapat di analisis dan disimpan. Data yang diambil dari media *online* berupa teks artikel dengan *keyword* yang diinput sebagai parameternya, kemudian di ekstrak ke dalam format *Excel (.CSV)* yang dilakukan dengan bantuan tool *Web Content Extractor (WCE)* dengan menggunakan teknik *Breadth-First Search*. Hasil yang didapat adalah data yang terstruktur berupa tabel dengan beberapa *field* yaitu no, hari/tanggal, waktu posting, judul, deskripsi, gambar, dan *link* halaman artikel. Berdasarkan analisis yang telah dilakukan, media *online* Detikcom lebih banyak menghasilkan data yang *relevan* dibandingkan dengan Liputan6 dan VivaNews, dan teknik *Breadth-First Search* sangat membantu dalam hal pencarian *record* dengan menelusuri URL utama hingga ke bagian terdalam *link*.

*Kata kunci: Web Scraping, Web Mining, Media Online, Breadth-First Search.*

## WEB SCRAPING ANALYSIS FOR NATURAL DISASTER USING BREADTH-FIRST SEARCH TECHNIQUE TOWARD THREE ONLINE MEDIA

### Abstract

*The current technologies enable people to work easily and quickly. The development of internet technology, make acquiring information is never been as easier as today. There is information retrieval technique that enables one to extract information provided from various websites, known as web scraping technique. The idea of this research is to collect information concerning natural disasters topic from three different online media (Detikcom, Liputan6, and VivaNews) by adopting web-scraping technique. Web scraping technique used to extract the unstructured data from the websites/online media and transform it into a form that analyzable. In this case, the data extracted from the online media are in a form of text article, with some keywords as the parameter. Moreover, the texts then converted into .CSV format, by using of Web Content Extractor (WCE) tool and adopting Breadth-First Search algorithm. The results of this research are structured data in form of table with corresponding fields, such as: number, day/date, posting time, title, description, image, and link of article page. From the analysis, Detikcom provide data that is more relevant to the research compared to Liputan6 and VivaNews. The Breadth-First Search algorithm is favorable in terms of finding the records by tracing the URL path.*

*Keywords : Web Scraping, Web Mining, Media Online, Breadth-First Search.*

## PENDAHULUAN

Kebiasaan yang mengglobal dari masyarakat saat ini adalah keseharian dengan internet. Data atau fakta apapun bisa diperoleh dengan akses internet tanpa batas waktu dan tempat, salah satunya media *online*. Dengan banyaknya media *online* atau situs web, masyarakat cenderung mencari informasi yang dibutuhkan melalui media *online* karena data yang ditampilkan *update* dan luas. Salah satunya tentang berita bencana alam.

Bencana alam merupakan sesuatu yang sangat sensitif dan butuh informasi yang cepat dan tepat. Melalui media *online*, akan ditemukan banyak sekali berita tentang bencana alam banjir, longsor, gempa bumi, dan lain-lain, baik yang baru terjadi ataupun yang sudah lampau. Data-data yang ditampilkan media *online* tersebut berskala besar dan belum terstruktur, sehingga akan kesulitan untuk menganalisis data-data lain yang berisi informasi serupa.

Berdasarkan data dari lembaga statistik BNPB (Badan Nasional Penanggulangan Bencana), 31,1% wilayah Indonesia merupakan daerah rawan banjir, 16,4% sering terjadi bencana tanah longsor, 20,2% sering terjadi puting beliung, 8,7% mengalami kekeringan, 12,2% sering terjadi kebakaran, dan sisanya 10,8% mencakup bencana tsunami, gempa bumi, letusan gunung api, kecelakaan transportasi, kebakaran hutan dan lain-lain. Dengan banyaknya wilayah di seluruh Indonesia yang sering terjadi musibah bencana alam, kemungkinan akan menyulitkan beberapa organisasi/kelompok untuk memperoleh data yang terstruktur dari media *online*. *Web scraping* (juga disebut *intelligent, automated, or autonomous agents*) hanya fokus pada cara mem-

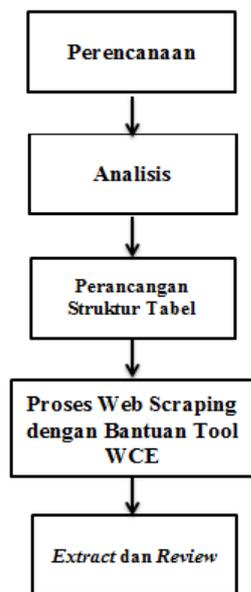
peroleh data melalui pengambilan dan ekstraksi data dengan ukuran data yang bervariasi.

Adapun Pembahasan pada penelitian ini memiliki batasan masalah, diantaranya: (1) Topik bencana alam yang dibahas hanya bencana alam banjir dan tanah longsor. (2) Media *online* yang digunakan untuk analisis *web scraping* adalah Liputan6, VivaNews, dan Detikcom. (3) Penelitian ini hanya mengekstraksi data berupa teks artikel dari media *online* ke dalam format *Excel (.CSV)* dengan bantuan *tools Web Content Extractor (WCE)*. (4) Data yang akan diambil hanya data pada tahun 2016. (5) Data yang diambil dari media *online* dibatasi maksimal 150 *record* untuk setiap *keyword* pada setiap media *online*, dikarenakan *tools WCE* yang masih *trial* atau tidak *free* yang membatasi maksimal 1 kali *running* adalah 150 *record*.

Adapun tujuan dari penelitian ini adalah melakukan analisis *web scraping* guna memperoleh data yang terstruktur dan mengumpulkan data-data yang tersebar agar informasi yang diambil lebih terfokus, sehingga memudahkan dalam melakukan pencarian data. Penelitian ini dilakukan agar data menjadi bentuk yang lebih bermakna dan dapat digunakan untuk kepentingan lain.

## METODE PENELITIAN

Metodologi atau alur pengembangan dari penelitian ini diterangkan oleh Gambar 1.



Gambar 1. Metode Penelitian

### ANALISA PERMASALAHAN

Pendefinisian masalah pada penelitian ini adalah kemudahan mendapatkan informasi yang tersebar di media *online* menjadi data yang terstruktur agar bisa lebih bermanfaat untuk kepentingan lain. Dan bagaimana peranan pemanfaatan metode *Breadth-First Search* pada analisis *web scraping*.

### ANALISA KEBUTUHAN

Pada tahap ini dilakukan analisis kebutuhan terhadap perangkat-perangkat yang akan digunakan untuk mendukung proses penelitian. Spesifikasi peralatan yang digunakan terdiri dari spesifikasi *hardware* dan *software*.

Perangkat keras (*hardware*) yang digunakan dalam pembuatan aplikasi ini adalah laptop dengan *processor* Intel(R) Atom(TM) CPU N2800 @ 1,86GHz 1,86 GHz, *memory* RAM 2.00 GB dan *harddisk* dengan kapasitas 500 GB.

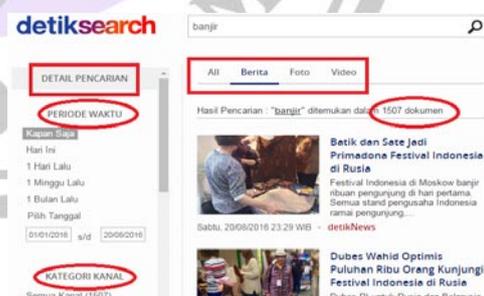
Sedangkan perangkat lunak (*software*) yang digunakan adalah sistem operasi windows 7 Ultimate 32-bit, *browser* Google Chrome, akses

internet, dan *tool Web Content Extractor* (WCE).

### ANALISA PEMILIHAN MEDIA ONLINE

Sebelum memutuskan untuk menggunakan media *online* Detikcom, Liputan6, dan VivaNews, yang harus diperhatikan adalah menganalisa apakah struktur halaman dari media *online* tersebut bisa melakukan pencarian dengan kata kunci, terhubung yang akan menjadi parameter perolehan data adalah kata kunci (*keyword*). Contoh kasus dalam hal ini adalah adanya *form* pencarian pada situs Detikcom, jika dilakukan penginputan *keyword* maka akan langsung menampilkan hasil pencarian, dan pada halaman tersebut juga terdapat *form* detail pencarian dimana kita bisa memilih periode waktu yang diinginkan dan kategori kanal untuk hasil pencarian. Sebagai contoh dapat dilihat pada gambar 2.

Selanjutnya yang harus diperhatikan adalah bagaimana hasil pencarian yang ditampilkan, apakah hasil pencarian yang ditampilkan sesuai dengan kata kunci yang diinput atau tidak.



Gambar 2. Contoh Struktur Halaman Media Online

### ANALISA PARAMETER MEDIA ONLINE

Setelah menganalisa media *online* yang akan digunakan, penelitian ini juga membutuhkan parameter sebagai acuan dalam perolehan data.

Yang menjadi parameter dalam pengambilan informasi dari media *online* adalah berdasarkan hasil *searching* pada media *online* dengan menggunakan *keyword* yang diinput.

Pemilihan *keyword* didasarkan pada topik yang diambil, yaitu banjir, tanah longsor, dan gempa bumi. Keterkaitan antara satu *keyword* dengan *keyword* yang lainnya sangat penting karena menyangkut hasil pencarian yang ditampilkan dan data yang akan diambil. Pemilihan *keyword* diawali dengan 1 kata *keyword* saja yang bersangkutan dengan topik utama, salah satunya “banjir”. *Keyword* “banjir” tersebut dipilih karena banjir adalah topik utama dari jenis bencana yang akan analisis.

Berikut adalah *keyword* yang akan digunakan sebagai parameter dalam pengambilan informasi dari media *online* dapat dilihat pada tabel 1

Tabel 1. Parameter yang Digunakan

No	Kategori	Keyword (Parameter)	
1	<b>Banjir</b>	• 1 kata	Banjir
		• 2 kata	Korban banjir
		• 3 kata	Korban banjir 2016
			Banjir jakarta 2016
			Bencana banjir indonesia
2	<b>Tanah Longsor</b>	• 1 kata	Tidak ada
		• 2 kata	Tanah longsor
			Longsor 2016
		• 3 kata	Korban longsor 2016
			Bencana longsor 2016
	• 4 kata	Longsor jawa barat 2016	

## GAMBARAN ANALISA WEB SCRAPING

Penelitian ini melakukan penerapan teknik *web scraping* dengan komponen-komponen utamanya seperti pada gambar 3.



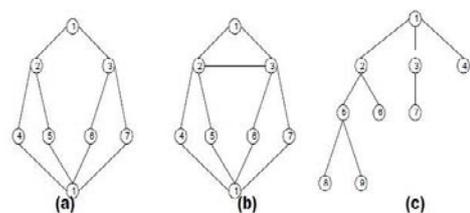
Gambar 3. Gambaran Analisa *Web Scraping*

Pada gambar 3 dapat dilihat bahwa langkah awal dimulai dengan menggunakan akses media *online* terlebih dahulu, melakukan input *keyword* pada media *online* sehingga halaman *website* menampilkan *list* artikel yang dimaksud oleh *keyword*. Kemudian melakukan proses *web scraping* pada *tool* WCE sehingga menghasilkan data dengan format tabel dan terakhir melakukan ekstraksi data menjadi data yang lebih terstruktur dan bermanfaat.

## METODE BREADTH-FIRST SEARCH

*Breadth-First Search (BFS)* adalah algoritma yang melakukan pencarian secara melebar yang mengunjungi simpul secara *pre-order*, yaitu mengunjungi suatu simpul kemudian mengunjungi semua simpul yang bertetangga dengan simpul tersebut terlebih dahulu. Selanjutnya, simpul yang belum dikunjungi dan bertetangga dengan simpul-simpul yang tadi dikunjungi, demikian seterusnya.

Berikut adalah contoh dari metode *Breadth-First Search* yang bisa dilihat pada gambar 4.



Gambar 4. Graf *Breadth-First Search*

Keterangan:

Gambar (a) *BFS*(1) : 1, 2, 3, 4, 5, 6, 7, 8

Gambar (b) *BFS*(1) : 1, 2, 3, 4, 5, 6, 7, 8

Gambar (c) *BFS*(1) : 1, 2, 3, 4, 5, 6, 7, 8, 9

Langkah-langkah penyelesaian menggunakan algoritma *Breadth-First Search (BFS)*:

1. Simpul ujung (akar) kedalam antrian.
2. Ambil simpul dari awal antrian, lalu cek apakah simpul merupakan solusi.
3. Jika simpul merupakan solusi, pencarian selesai dan hasil dikembalikan.
4. Jika simpul bukan solusi, masukkan seluruh simpul yang bertetangga dengan simpul tersebut (simpul anak) kedalam antrian.
5. Jika antrian kosong dan setiap simpul sudah dicek, pencarian selesai dan mengembalikan hasil "solusi tidak ditemukan".

Pada metode *breadth first search*, semua *node* pada level *n* akan dikunjungi terlebih dahulu sebelum mengunjungi *node* pada level *n+1*. Pencarian dimulai dari *node* akar terus ke level 1 dari kiri ke kanan, kemudian berpindah ke level berikutnya demikian pula dari kiri ke kanan hingga ditemukan solusinya.

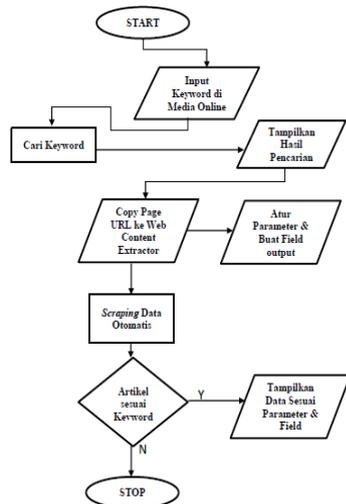
Terdapat tiga langkah yang dilakukan pada tahap *crawling* ketika mengunjungi artikel, yaitu menandai bahwa suatu artikel telah dikunjungi, mengenali link yang terdapat pada artikel tersebut, kemudian isinya didaftarkan pada daftar indeks. Pada

akhirnya, *web scraping tools* akan menampilkan file yang paling banyak berkaitan dengan kata kunci.

## FLOWCHART WEB SCRAPING

Flowchart adalah diagram alur atau logika pemrosesan suatu program atau sistem. Flowchart *web scraping* pada penelitian ini didasarkan pada *tools* yang digunakan, yaitu Web Content Extractor. Pada gambar 5 dapat dilihat bahwa proses awal diawali dengan *Start*. Langkah pertama yang dilakukan adalah menginput *keyword* yang diinginkan pada media *online* di *browser* Chrome. Dengan demikian *website* tersebut akan memproses pencarian artikel sesuai *keyword* dan menampilkan semua artikel yang terkait dengan *keyword*. Setiap media *online* menampilkan hasil yang berbeda-beda dengan input *keyword* yang sama. Ada yang menampilkan pencarian yang *relevan*, dan ada juga yang secara umum berdasarkan tag *keyword* yang diinput.

Selanjutnya input page URL ke dalam *tool* Web Content Extractor. Pada *tools* tersebut kita akan mengatur parameter dan buat *field* yang dibutuhkan untuk output. Setelah itu *tool* WCE akan melakukan *scraping* secara otomatis pada artikel yang sesuai *keyword*, jika proses benar maka akan dihasilkan data sesuai *field* dan parameter. Jika tidak, maka proses berhenti dan proses *running* selesai dilakukan.



Gambar 5. Flowchart Web Scraping

3.	Waktu Posting	Menunjukkan waktu (jam) kapan artikel di posting
4.	Judul	Berisi judul dari artikel
5.	Deskripsi	Berisi keterangan atau isi dari artikel
6.	Gambar	Berisi gambar utama dari artikel
7.	Link	Merupakan alamat URL dari halaman artikel

## HASIL DAN PEMBAHASAN

### PERANCANGAN STRUKTUR TABEL DATA

Perancangan tabel data ini dapat dilihat pada tabel 2. Tabel data tersebut dibutuhkan untuk mengetahui apa saja data yang diinginkan dan agar data yang didapat lebih terstruktur sesuai dengan struktur tabel yang sudah dirancang.

Tabel 2. Struktur Tabel Data

No	Nama Tabel	Keterangan
1.	No	Berisi nomor atau ID
2.	Hari/Tanggal	Merupakan hari/tanggal postingan

Hasil dari penelitian ini adalah data yang terstruktur yaitu berupa tabel dengan format Excel (.CSV) seperti pada gambar 6.

Gambar 6 merupakan tampilan hasil penyortiran data dari hasil proses web scraping, dimana proses penyortiran tersebut dimaksudkan untuk mengambil data-data yang relevan dan memisahkannya dari data-data yang tidak relevan. Proses penyortiran tersebut disebut juga dengan data extraction.

No	Jenis Berita	Media Online	Hari/Tanggal	Waktu Posting	Judul	Deskripsi	Gambar	Link
1	Banjir	Detikcom	Sabtu 02 Jul 2016	01:41 WIB	Tanggap Jebol, Ratusan	Bandung - Tanggul se	e9b0d1f2-1f9e-	<a href="http://news.detik.com">http://news.detik.com</a>
2	Banjir	Detikcom	Jumat 01 Jul 2016	17:27 WIB	Jalan Nasional Surabaya	Jakarta - Hujan deras	8a4cef77-e434-	<a href="http://news.detik.com">http://news.detik.com</a>
3	Banjir	Detikcom	Jumat 01 Jul 2016	14:46 WIB	Bupati Perintahkan Din	Pasuruan - Selain me	704a9a85-55b8-	<a href="http://news.detik.com">http://news.detik.com</a>
4	Banjir	Detikcom	Kamis 30 Jun 2016	14:14 WIB	Tiga Warga Pasuruan	Pasuruan - Derasnya	33ab5b60-5868-	<a href="http://news.detik.com">http://news.detik.com</a>
5	Banjir	Detikcom	Minggu 26 Jun 2016	13:48 WIB	Anggota DPR Minta M	Jakarta - Bencana ala	a1e44ea3-921d-	<a href="http://news.detik.com">http://news.detik.com</a>
6	Banjir	Detikcom	Sabtu 25 Jun 2016	07:45 WIB	Tanggap Darurat Benc	Jakarta - Bencana bar	60b7eef4-82f4-	<a href="http://news.detik.com">http://news.detik.com</a>
7	Banjir	Detikcom	Sabtu 25 Jun 2016	05:14 WIB	Operasi SAR di Purwo	Jakarta - Setelah 7 ha	2380c8c7-93bf-	<a href="http://news.detik.com">http://news.detik.com</a>
8	Banjir	Detikcom	Rabu 22 Jun 2016	19:11 WIB	Didatangi Ganjar, Korb	Banyumas - Gubernur	df35feb3-103d-	<a href="http://news.detik.com">http://news.detik.com</a>
9	Banjir	Detikcom	Rabu 22 Jun 2016	15:14 WIB	Pemudik Diimbau Was	Jakarta - Banjir rob d	237c0d89-34bf-	<a href="http://news.detik.com">http://news.detik.com</a>
10	Banjir	Detikcom	Rabu 22 Jun 2016	07:34 WIB	Banjir dan Longsor Ter	Jakarta - Hujan lebat	4c993ecc-1435-	<a href="http://news.detik.com">http://news.detik.com</a>
11	Banjir	Detikcom	Rabu 22 Jun 2016	00:02 WIB	Kerugian Akibat Banjir	Banyumas - Kerugian	b47d1b81-fa13-	<a href="http://news.detik.com">http://news.detik.com</a>
12	Banjir	Detikcom	Selasa 21 Jun 2016	20:03 WIB	Warga Korban Longsor	Semarang - Sejumlah	0a92982f-0935-	<a href="http://news.detik.com">http://news.detik.com</a>
13	Banjir	Detikcom	Selasa 21 Jun 2016	16:29 WIB	Sehari Dibuka, Posko	Surabaya - Baru seha	50daead-e2c7-	<a href="http://news.detik.com">http://news.detik.com</a>
14	Banjir	Detikcom	Selasa 21 Jun 2016	15:06 WIB	Begini Kerusakan Paral	Sangihe - Bencana ba	2667ddf9-e364-	<a href="http://news.detik.com">http://news.detik.com</a>
15	Banjir	Detikcom	Selasa 21 Jun 2016	13:26 WIB	Banjir dan Longsor Ter	Sangihe - Bencana ba	9591f8e6-bd37-	<a href="http://news.detik.com">http://news.detik.com</a>
16	Banjir	Detikcom	Selasa 21 Jun 2016	09:39 WIB	Mendikbud Tinjau Banj	Solo - Mendikbud An	209799d1-b3f6-	<a href="http://news.detik.com">http://news.detik.com</a>
17	Banjir	Detikcom	Selasa 21 Jun 2016	01:12 WIB	Pemerintah Imbau Dan	Jakarta - Bencana bar	f33da853-e9c3-	<a href="http://news.detik.com">http://news.detik.com</a>
18	Banjir	Detikcom	Senin 20 Jun 2016	15:41 WIB	Jawa Tengah Banjir da	Surabaya - Wali Kota	bb834c8c-d777-	<a href="http://news.detik.com">http://news.detik.com</a>
19	Banjir	Detikcom	Senin 20 Jun 2016	13:32 WIB	600 warga Masih teris	Jakarta - Pasca banjir	5cf75ea8-e1ff-	<a href="http://news.detik.com">http://news.detik.com</a>

Gambar 6. Tampilan Hasil Analisis Web Scraping

## HASIL PERBANDINGAN PEROLEHAN DATA

Setelah melakukan data extraction dengan memilih data-data yang relevan, maka hasil data extraction dapat dilihat pada tabel 3.

Berdasarkan tabel 3 dapat dilihat bahwa masing-masing bencana memiliki 5 parameter yang diambil dari keyword yang diinput pada media online. Data yang diambil adalah data yang relevan. Data yang sudah diambil berdasarkan keyword sebelumnya pada 1 media online, tidak akan diambil lagi jika masih ada pada hasil web scraping dengan keyword yang berbeda. Media yang digunakan ada 3, yaitu Detikcom, Liputan6, dan VivaNews. Jika dilihat dari kolom jumlah hasil scraping, media online detikcom memiliki jumlah terbanyak yang menampilkan hasil pencarian pada situsnya.

banyak menghasilkan data yang relevan dibandingkan dengan bencana longsor. Dan total jumlah keseluruhan data yang relevan adalah 922 record. Sedangkan data yang tidak relevan untuk bencana banjir berdasarkan 3 media online tersebut berjumlah 497 record, dan 640 record untuk bencana longsor. Sangat terlihat bahwa bencana longsor memiliki jumlah data yang lebih banyak tidak relevan dibandingkan bencana banjir. Total keseluruhan data yang relevan lebih sedikit, yaitu 922 record dari pada data yang tidak relevan, yaitu 1137 record.

### Hasil Perbandingan Analisis Web Scraping Terhadap Media Online

Berdasarkan hasil analisa web scraping terhadap 3 media online, detikcom, liputan6, dan vivanews, maka diperoleh tabel hasil perbandingan analisis web scraping terhadap 3 media online seperti pada tabel 4.

Tabel 3. Hasil Perbandingan Perolehan Data

No	Jenis Bencana	Parameter	Jumlah Hasil Scraping (record)			Hasil Data Extraction (record)						Jumlah Data yang Relevan per Jenis	Jumlah Data yang Tidak Relevan per Jenis
			Detik	Liputan6	VivaNews	Data yang relevan			Data yang tidak relevan				
						Detik	Liputan6	VivaNews	Detik	Liputan6	VivaNews		
1	Banjir	Banjir	150	150	86	108	73	49	42	77	37	621	497
		Korban Banjir	150	21	150	108	0	32	42	21	118		
		Korban Banjir 2016	150	20	0	108	12	0	42	8	0		
		Banjir Jakarta 2016	150	20	43	83	10	31	67	10	12		
		Bencana Banjir di Indonesia	8	20	0	0	7	0	8	13	0		
2	Longsor	Tanah Longsor	146	56	150	31	25	4	115	31	146	301	640
		Longsor 2016	150	11	11	37	8	2	113	3	9		
		Korban Longsor 2016	150	20	0	94	17	0	56	3	0		
		Bencana Longsor 2016	150	20	0	63	5	0	87	15	0		
		Longsor Jawa Barat 2016	57	20	0	12	3	0	45	17	0		
Jumlah						644	160	118	617	198	322		
Jumlah Total											922	1137	

Setelah dilakukan ekstraksi data, maka jumlah data yang relevan terkait bencana banjir pada 3 media online tersebut ada 621 record. Sedangkan untuk bencana longsor, jumlah data yang relevan berdasarkan 3 media online tersebut ada 301 record. Jadi, jenis bencana banjir lebih

Tabel 4. Tabel Perbandingan Terhadap Media Online

No	Media Online	Keterangan
1	Detikcom	<ul style="list-style-type: none"> <li>Struktur halaman pencarian pada detikcom lebih spesifik, yakni terdapatnya detail pencarian untuk pemilihan berdasarkan tanggal dan kategori kanal, juga menampilkan jumlah artikel untuk hasil pencarian tersebut.</li> <li>Media <i>online</i> detikcom memiliki jumlah pencarian terbanyak pada situsny, baik dengan 1 <i>keyword</i>, maupun 2 dan 3 <i>keyword</i> dibandingkan 2 media lainnya.</li> <li>Tingkat <i>ke-relevan-an</i> data pada detikcom lebih banyak dibandingkan dengan 2 media lainnya.</li> <li>Jumlah data yang <i>relevan</i> lebih banyak daripada 2 media lainnya, yaitu sebanyak 644 <i>record</i>, hal itu didasarkan pada tingkat <i>ke-relevan-an</i> hasil pencarian datanya.</li> </ul>
2	Liputan 6	<ul style="list-style-type: none"> <li>Struktur halaman pencarian pada liputan6 kurang bagus, karena tidak terdapatnya detail pencarian untuk pemilihan berdasarkan tanggal dan tidak ditampilkannya jumlah artikel pada hasil pencarian.</li> <li>Media <i>online</i> liputan6 memiliki jumlah pencarian terbanyak ke-2 pada situsny setelah detikcom, baik dengan 1 <i>keyword</i>, maupun 2, 3, dan 4 <i>keyword</i>.</li> <li>Tingkat <i>ke-relevan-an</i> datanya lebih sedikit jika dibandingkan dengan detikcom.</li> <li>Jumlah data yang <i>relevan</i> lebih sedikit daripada detikcom, yaitu 160 <i>record</i>.</li> </ul>
3	VivaNews	<ul style="list-style-type: none"> <li>Struktur halaman pencarian pada vivanews lebih bagus daripada liputan6, karena ditampilkannya jumlah artikel pada hasil pencarian, namun tidak ada detail pencarian untuk pemilihan berdasarkan tanggal.</li> <li>Media <i>online</i> vivanews memiliki jumlah pencarian paling sedikit pada situsny dibandingkan 2 media lainnya. Vivanews hanya menampilkan pencarian berdasarkan <i>keyword</i> yang umum atau terdiri dari 2 kata. Sedangkan untuk <i>keyword</i> yang lebih spesifik seperti 3 atau 4 kata, vivanews tidak memiliki hasil pencarian.</li> <li>Tingkat <i>ke-relevan-an</i> datanya juga lebih sedikit jika dibandingkan dengan 2 media lainnya.</li> <li>Jumlah data yang <i>relevan</i> lebih sedikit daripada detikcom dan liputan6, dikarenakan hasil pencarian pada situsny dan tingkat <i>ke-relevan-an</i> datanya.</li> </ul>

## KESIMPULAN DAN SARAN

### KESIMPULAN

Berdasarkan hasil analisa web scraping terhadap media online,

parameter pada media online sangat berpengaruh terhadap proses web scraping. Hal itu dikarenakan keharusan dalam menginput URL pada tool WCE sebelum melakukan proses scraping. Data yang di dapat dari hasil web scraping pada masing-masing media juga sangat berbeda dari tingkat ke-relevan-an dan jumlah data yang didapat. Dari ketiga media online tersebut, situs Detikcom lebih banyak menghasilkan data yang relevan berdasarkan keyword-nya dibandingkan dengan Liputan6 dan VivaNews.

Sedangkan untuk metode yang digunakan oleh tool WCE, yaitu metode Breadth-First Search sangat membantu dalam hal crawling pada proses scraping data dari website. Karena metode ini membaca semua link pada tingkatan yang sama terlebih dahulu, kemudian menelusuri bagian-bagian terdalam link untuk menemukan data.

### SARAN

Hasil dari penelitian ini dapat digunakan untuk kepentingan analisis terkait bencana alam terhadap media online, seperti analisis Big Data dimana kebutuhan data yang besar dan akurat sangat diperlukan. Selain itu juga disarankan untuk melakukan otomatisasi pada penyortiran data yang relevan dan tidak relevan agar lebih mempersingkat waktu dan meningkatkan pemilihan ke-relevan-an data. Diharapkan penelitian ini akan berguna kedepannya bagi pembaca maupun penulis.

### DAFTAR PUSTAKA

- [1] A. Josi, L.A. Abdillah, Suryayusra. *Penerapan Teknik Web Scraping Pada Mesin Pencari Artikel Ilmiah*. Program Studi Teknik Informatika dan Sistem

- Informasi, Fakultas Ilmu Komputer,  
Universitas Bina Darma.
- [2] Cynthia Kustanto, Ratna Mutia, Pocut  
Viqarunnisa. *Penerapan Algoritma  
Breadth-first Search dan Depth-first  
Search Pada FTP Search Engine for ITB  
Network*. Departemen Teknik  
Informatika, Institut Teknologi Bandung.
- [3] Govind Murari Upadhyay, Kanika  
Dhingra, 2013. *Web Content Mining: Its  
Techniques and Uses*. International  
Journal of Advanced Research in  
Computer Science and Software  
Engineering. Volume 3, Issue 11, Hal  
610-613
- [4] Mardi Siswo Utomo, 2013. *Web  
Scraping pada Situs Wikipedia  
menggunakan Metode Ekspresi Regular*.  
Jurnal Teknologi Informasi DINAMIK.  
Volume 18, No.2: 153-160.
- [5] Mrs.C.Menaka M.C.A., M.Phil.,  
Dr.N.Nagadeepa. M.Sc., M.Phil.,  
M.C.A.,Ph.D, 2014. *A Survey of Web  
Content Mining Tools and Future  
Aspects*. International Journal of  
Advanced Research in Computer Science  
Engineering and Information  
Technology. Volume 3, Issue 1, Hal 375-  
385.

