Ekstraksi Informasi Terstruktur Profil Pengguna Website Iklan Baris

Nuri Cahyono[1]
Program Studi Informatika, Fakultas Ilmu Komputer
Universitas Amikom Yogyakarta
Jl. Padjajaran, Ring Road Utara,55281, Daerah Istimewa Yogyakarta, Indonesia
Email: [1]nuricahyono@amikom.ac.id

Abstract. Extraction of Structured Information on Classified Ads Website User Profiles. *The large amount of user data published on online buying and selling sites provides benefits for research and digital marketing. Data extraction was a method for obtaining data from publicly published website content. The first step was to determine the website category that was needed, then determined the attributes to be used as a reference in compiling the data extraction tags. The next step was to identify the tags that were taken based on the tag elements of the website that matched these attributes. Elements to compile tag extraction included CSS Selector, HTML Tag and Xpath. Based on this, a data extraction code was created with four iterations based on categories. The test was done by calculating the accuracy to find out the complete amount of extracted data. From a total of 16,000 data extracted in this test, the accuracy rate was 99.0625%.*
Keywords: *Data Extraction, Web Scrapping, Classified Ads*

Abstrak. *Perkembangan situs jual beli online berdampak pada banyaknya data pengguna yang dipublikasikan secara online. Profil pengguna situs web memiliki banyak manfaat baik untuk penelitian maupun untuk tujuan dalam pemasaran digital. Ekstraksi data adalah mekanisme untuk mendapatkan data dari konten situs web yang disajikan secara umum. Langkah pertama adalah menentukan kategori website kemudian menentukan atribut yang akan digunakan sebagai acuan dalam menyusun tag ekstraksi data yang diambil berdasarkan elemen tag dari website yang sesuai dengan atribut tersebut. Elemen tag yang diambil untuk menyusun tag ekstraksi antara lain CSS Selector, HTML Tag dan Xpath, kemudian dibuat skenario ekstraksi data dengan skenario empat kasus berdasarkan kategori yang telah ditentukan. Pengujian dilakukan dengan menghitung akurasi untuk mengetahui jumlah data yang berhasil di dapatkan secara lengkap. Dari total 16000 data dari hasil ekstraksi, dalam pengujian ini menghasilkan tingkat akurasi 99.0625%.*
Kata Kunci: *Ekstraksi Data, Web Scrapping, Iklan Baris*

1. Pendahuluan

Situs web jualan daring sudah menjadi bagian yang tidak terpisahkan dari era teknologi informasi seperti sekarang ini, berbagai macam mekanisme dan gaya dari situs web yang ada di indonesia sangat mempengaruhi dari layanan yang disediakan. Layanan pada situs web jualan daring tidak sebatas hanya dari transaksi jual-beli saja, tetapi juga berbagai proses yang ada di dalamnya [1]. Setiap situs web jualan biasanya memiliki karakteristik sendiri dari prosesnya ada situs web dengan mekanisme semua transaksi harus menggunakan sistem sehingga tidak ada data pengguna yang dipublikasikan, tetapi ada juga yang menggunakan mekanisme transaksi langsung antar pengguna hal ini yang biasanya terdapat di situs web iklan baris [2]. Pada iklan baris memiliki data profil pengguna yang bersifat umum sehingga bisa di akses dan dimanfaatkan oleh semua orang, data yang ada bervariasi tergantung fitur yang disediakan oleh situs web iklan baris tersebut. Ekstraksi data merupakan proses pengambilan konten dari terstruktur atau semi terstruktur dari suatu situs web, umumnya dari halaman situs web yang memiliki struktur HTML atau XHTML dengan cara menganalisis konten yang dibutuhkan sebelum nantinya hasil dari ekstraksi data tersebut akan dimanfaatkan untuk berbagai

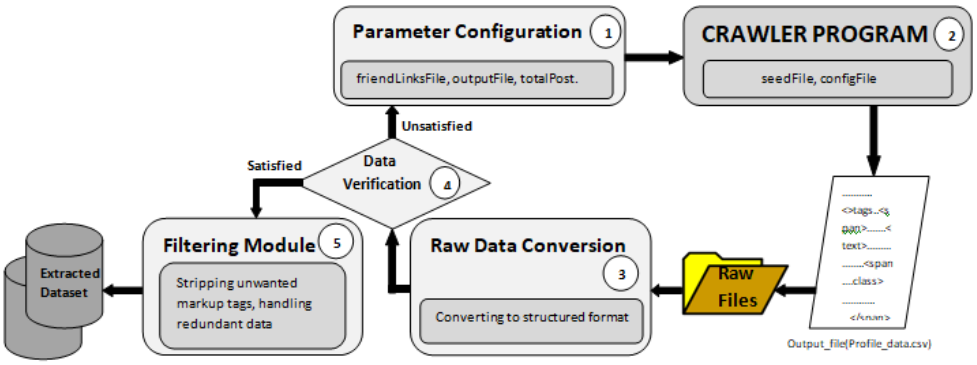
kepentingan. Bahasa utama yang digunakan untuk ekstraksi data yaitu bahasa markup, bahasa markup sendiri merupakan bahasa komputer yang menggunakan tag untuk melakukan definisi terhadap konten situs web [3].

Ada berbagai pendekatan yang dapat gunakan untuk melakukan ekstraksi data pada situs web salah satunya dengan memanfaatkan metode adaptif yaitu metode untuk melakukan ekstrak data dari target setelah halaman web diperbaharui berdasarkan aturan pencocokan fitur teks dan fitur tag html dari suatu halaman situs web [4]. Fitur dengan memanfaatkan tag sangat cocok digunakan tanpa harus menunggu atau mendapatkan layanan ekstraksi yang tersedia. Setiap halaman situs web memiliki gaya dan desain yang berbeda dalam pembuatannya biasanya struktur yang dibuat akan semakin rumit. Tetapi pada umumnya memiliki satu kesamaan yaitu bagian tag untuk menyajikan informasi yang dibuat. Seperti penelitian yang dilakukan oleh Carrey yang mempelajari struktur situs web berdasarkan *paragraph tag* yang dapat menjadi solusi tentang perbedaan gaya dan desain pada situs web yang hadir selama ini [5]. Dari hasil penelitian tersebut dihasilkan nilai *precision* tertinggi 94.3% dan nilai *precision* terendah 83.0%, sedangkan untuk nilai *recall* tertinggi 97.4% dan nilai *recall* tersendah 92.9%.

Dalam penelitian ini penulis akan melakukan penelitian ekstraksi data web iklan baris dengan menggunakan pendekatan elemen tag yang terdiri dari CSS Selector, HTML Tag dan Xpath, dengan menggunakan skenario empat kasus ekstraksi data. Hasil konten situs web yang berupa profil pengguna akan dilakukan evaluasi untuk mengetahui akurasi berdasarkan hasil ektraksi data yang sesuai dengan data yang tidak lengkap.

2. Tinjauan Pustaka

Dalam perkembangan data daring yang cukup pesat sehingga banyak dilakukan proses ekstraksi data untuk berbagai keperluan dan dari banyak situs web yang menyajikan data secara umum sehingga bisa dimanfaatkan oleh banyak kalangan. Seperti penelitian yang dilakukan oleh Wani, dkk. mengatakan bahwa metode ekstraksi data memiliki banyak tantangan terutama bagi situs web yang tidak menyediakan API secara independen, sehingga metode berdasarkan elemen memiliki peran penting dan memudahkan dalam melakukan penelitian ekstraksi data publik. Gambaran langkah ekstraksi data yang dilakukan dapat dilihat pada Gambar 1. Proses ekstraksi dilakukan berdasarkan informasi pribadi yang diungkapkan oleh penulis secara umum [6].



Gambar 1. ***Data Collection Framework***

Berbagai penelitian terdahulu tentang ektraksi data seperti yang di sajikan dalam Tabel 1. Mehak dalam penelitiannya, melakukan ekstraksi data pada informasi penawaran produk dengan skenario lima kasus yang memiliki kerangka situs web HTML dengan memanfaatkan Selenium [7]. Sementara itu Surahman melakukan ekstraksi data dari produk *marketplace* yang ada di Indonesia [8]. Penelitian yang dilakukan oleh Polidoro melakukan pengujian hasil *scrapping* web dalam bidang *e-commerce*, yaitu survei harga konsumen sebuah produk elektronik dan tiket pesawat [9]. Penelitian Akbar, dkk. menjelaskan bagaimana melakukan ekstraksi halaman web berdasarkan *column-row wise* yang akan diinputkan ke dalam sebuah

database [10]. Penelitian tentang ekstraksi halaman web untuk mengetahui keuntungan dalam penggunaan, proses ekstraksi yang dilakukan berdasarkan halaman situs web yang diterapkan pada *mobile app* [11]. Ekstraksi data konten situs web dengan memanfaat struktur pendekatan DOM seperti penelitian yang dilakukan oleh Yu, melakukan ekstraksi dari berbagai halaman situs web dengan cara menganalisa struktur DOM dari situs web tersebut [12].

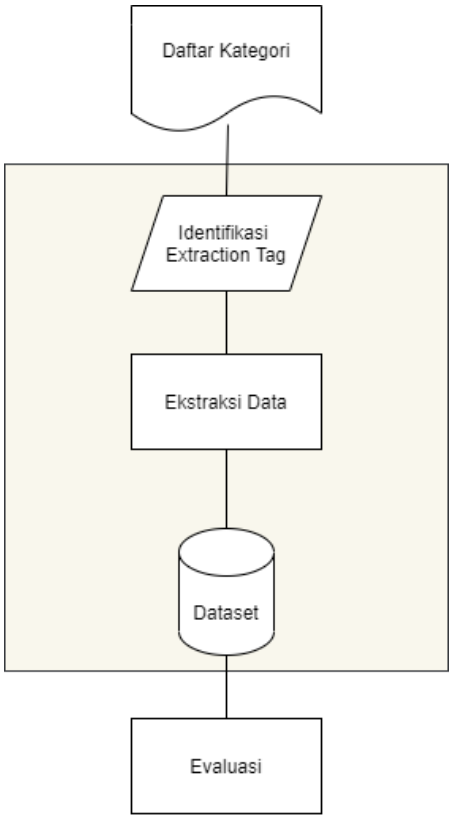
Tabel 1. Penelitian Terdahulu

Penulis	Judul	Pembahasan	Hasil
S. Mehak, R. Zafar, S. Aslam and S. M. Bhatti	Exploiting Filtering approach with Web Scrapping for Smart Online Shopping : Penny Wise: A wise Tool for Online Shopping	Ekstraksi data pada informasi penawaran produk dengan skenario lima kasus yang memiliki kerangka situs web html dengan memanfaatkan selenium dalam proses ekstraksi datanya	Memperoleh nilai akurasi sebesar 93%
Surahman, A. F. Octaviansyah dam D. Darwis	Ekstraksi Data Produk E-*Marketplace* Sebagai Strategi Pengolahan Segmentasi Pasar Menggunakan Web Crawler	Melakukan ekstraksi data dari produk *marketplace* yang ada di Indonesia dengan memanfaatkan teknologi web crawler dan pengujian menggunakan sebanyak 250 responden	Hasil kesuksesan informasi segmentasi sebesar 79%
F. Polidoro, R. Giannini, R. L. Conte, S. Mosca and F. Rosseti	Web Scrapping Techniques to collect data on consumer electronics and airfares for italian HICP compilation	Penelitian ini memberikan gambaran tentang pengujian teknik web scrapping dalam bidang survei harga konsumen dengan referensi khusus untuk produk elektronik (barang) dan tiket pesawat (jasa).	Merekomendasikan web scrapping untuk ektraksi data dan kombinasikan dengan statistic untuk evaluasi hasil
M. Akbar and A. Wibowo	Ekstraksi tabel html bentuk column-row wise ke dalam basis data.	Menjelaskan bagaimana melakukan ekstraksi halaman web berdasarkan *column-row wise* yang akan diinputkan ke dalam sebuah database, dengan memastikan sebelumnya keterhubungan antara atribut dan tabel tidak hilang	Kategori yang digunakan memiliki kompleksitas tinggi dengan nilai kompleksitas 12
A. Sasongko	Integrasi Data Website Student.BSI.AC.ID Untuk Mobile Infokampus Berbasis Android Menggunakan Ekstraksi HTML	Penelitian tentang ekstraksi halaman web untuk mengetahui keuntungan dalam penggunaan, proses ekstraksi yang dilakukan berdasarkan halaman situs web yang diterapkan pada mobile app	Kesimpulan mulai dari ekstraksi halaman situs web dalam hal konsumsi bandwith 91.54% atau rata-rata perhalaman sebesar 16.68, tetapi proses akses perhalaman 1.75 yang artinya lebih panjang dari pada browser sebesar 117.68%
X. Yu and Z. Jin	Web Content Information Extraction Based on DOM Tree and Statistical Information	Melakukan ekstraksi konten dari berbagai halaman situs web dengan cara menganalisa struktur DOM.	Precision 98,21% Recall 97.90%

3. Metodologi Penelitian

Halaman situs web untuk penjualan daring memiliki banyak variasi untuk saat ini mulai dari konsep toko daring, katalog produk, *marketplace* hingga iklan baris. Dalam penelitian mengambil objek berdasarkan situs web penjualan daring iklan baris. Iklan baris menggunakan mekanisme proses dalam transaksi yaitu langsung dengan menghubungi kontak dari pengguna yang melakukan *posting* produk sehingga memiliki data pengguna yang disajikan secara terbuka dan dapat diakses secara umum.

Penelitian ekstraksi data ini merupakan sebuah penelitian eksperimental seperti yang di tunjukan pada Gambar 2. Tahap awal dimulai dengan menentukan objek situs web iklan baris beserta kategori pada webiste tersebut yang akan dijadikan sebagai target dalam ekstraksi data. Terdapat empat atribut dalam penelitian ini yaitu nama (N), kategori (K), kontak (H) dan alamat (A) seperti pada Tabel 1. Skenario berikutnya yaitu dengan melakukan identifikasi *extraction tag* yang ada berdasarkan atribut yang sudah ditentukan untuk digunakan menyusun kode ekstraksi data.



Gambar 2. Metodologi Penelitian

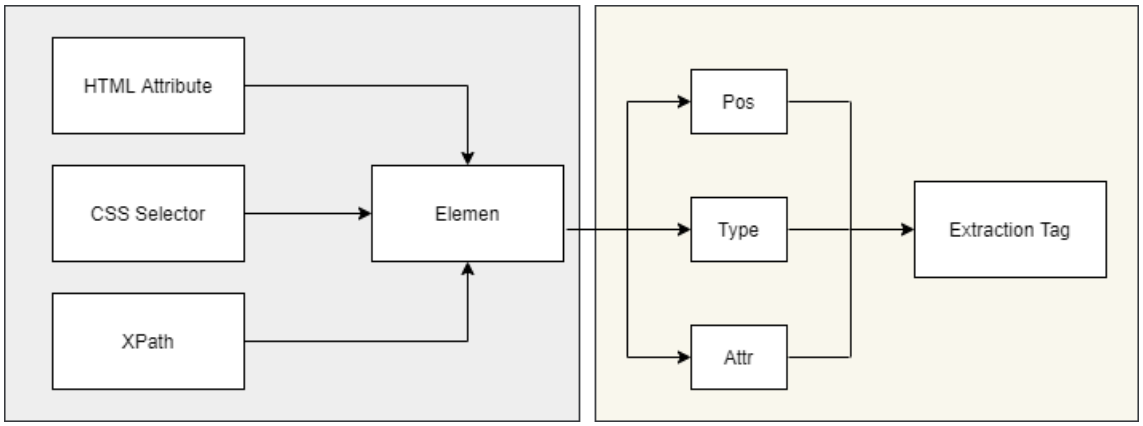
Proses ekstraksi data dilakukan secara otomatis berdasarkan *tag & selector* yang telah disusun sehingga sesuai dengan atribut yang ditentukan sebelumnya. Dalam skenario ekstraksi terdapat empat kali percobaan dengan kategori yang berbeda, dengan masing-masing hasil nya akan disimpan ke dalam *database*. *Dataset* yang telah dihasilkan kemudian akan dilakukan evaluasi apakah hasil ekstraksi sudah sesuai dengan skenario yang diharapkan dalam penelitian ini.

3.1. Extraction Tag

Untuk memperoleh kelengkapan data dalam proses ektraksi data dengan teknik web *scrapping* ada beberapa hal yang bisa digunakan XPath Selector dan CSS Selector yang keduanya tergabung menjadi satu dalam HTML atribut [13]: (1) HTML Attribute, berdasarkan atribut html yang diidentifikasi dari situs web target kemudian diambil sebagai elemen yang nantinya akan disusun kedalam *extraction tag,* (2) CSS Selector, berdasarkan CSS selector yang diidentifikasi

dari situs web taget kemudian diambil sebagai elemen yang nantinya akan disusun kedalam *extraction tag,* (3) Xpath, berdasarkan Xpath yang telah diidentifikasi dari situs web target diambil sebagai elemen yang nantinya akan disusun kedalam *extraction tag.*

Dalam proses identifikasi dan penyususan elemen menjadi *extraction tag* yang sebelumnya sudah disesuaikan dengan kebutuhan atribut untuk penelitian ekstraksi data ini lebih jelasnya seperti yang tercantum pada Gambar 3. Proses di mulai dengan melakukan identifikasi situs web target atau sumber untuk di cocokan dengan salah satu atau ketiga tag yang telah di definisikan yaitu HTML Attribute, CSS Selector dan Xpath. Dari tag tersebut dicek bagian yang menampung empat atribut nama (N). kategori (K), kontak (H) dan alamat (A). Setelah di temukan kecocokan kemudian tag yang sesuai tersebut diambil untuk dimasukkan ke dalam elemen. Tidak semua dari tag akan diambil untuk dimasukkan ke dalam elemen, bisa salah satu atau ketiganya sekaligus yang akan diambil hal itu disesuaikan dengan kebutuhan data yang akan diekstraksi dan juga struktur dari situs web tersebut menggunakan metode tag yang seperti apa.



Gambar 3. *Extraction Tag*

Elemen yang telah berhasil diidentifikasi kemudian disusun menjadi elemen yang di dalamnya terdapat tiga parameter yang harus dilengkapi agar bisa menjadi sebuah model untuk melakukan ekstraksi data. Ketiga parameter tersebut yaitu *pos*, *type* dan *attr* dari masing masing elemen memiliki ketiganya yang diambil dari identifikasi tag yang terdapat dalam struktur situs web. *Extraction tag* berhasil disusun akan digunakan kedalam empat tahapan skenario ekstraksi data, dengan satu tag tadi akan dilakukan perulangan dalam proses ekstraksi terhadap empat kasus yang didefiniskan pada pembahasan Tabel 2. Untuk melihat lebih detail tentang kode *extrasi tag* yang akan disusun sebagai berikut ini Gambar 4.yang memberikan gambaran tentang elemen dengan tiga parameter.

```
1 Tag POS=CONTOH1 TYPE=CONTOH1 ATTR=CONTOH1
2 Tag POS=CONTOH2 TYPE=CONTOH2 ATTR=CONTOH2
3 Tag POS=CONTOH3 TYPE=CONTOH3 ATTR=CONTOH3
4 Tag POS=CONTOH4 TYPE=CONTOH3 ATTR=CONTOH4
```

Gambar 4. Contoh Kode

Kode tag ekstraksi data disusun berdasarkan tag yang ada di struktur situs web di ambil tiga parameter yang disesuaikan dengan atribut yang akan diekstraksi. Pada Gambar 4 menjelaskan bagaimana hasil identifikasi tag disusun menjadi kode ekstraksi data: (1) Pada baris pertama terdapat parameter *pos*, *type* dan *attr* yang berisi contoh 1 merupakan hasil identifikasi tag berdasarkan atribut pertama, (2) Pada baris kedua terdapat parameter *pos*, *type* dan *attr* yang berisi contoh 2 merupakan hasil dari identifikasi tag berdasarkan atribut kedua, (3) Pada baris ketiga terdapat parameter *pos*, *type* dan *attr* yang berisi contoh 2 merupakan hasil

dari identifikasi tag berdasarkan atribut ketiga, (4) Pada baris keempat terdapat parameter *pos*, *type* dan *attr* yang berisi contoh 2 merupakan hasil dari identifikasi tag berdasarkan atribut keempat.

Susunan kode diatas ditambahkan beberapa kode sesuai dengan kebutuhan atau di sesuaikan dengan struktur situs web yang ada. Selanjutnya akan digunakan dalam proses ekstraksi data.

3.2. Evaluasi Hasil Ektraksi Data

Dalam tahap ini dilakukan evaluasi berdasarkan hasil dari proses ekstraksi data yang dilakukan, dengan tujuan untuk mengetahui jumlah data yang memilik atribut lengkap dari total semua data hasil ekstraksi data. Perhitungan evaluasi dilakukan secara manual yaitu dari jumlah keseluruhan data yang diekstraksi akan di kelompokan menjadi dua yaitu data hasil extraksi dengan attribute lengkap diberikan skor (0) dan data hasil ekstraksi yang memiliki attribute tidak lengkap diberikan atribut (1). Tingkat keberhasilan keseluruhan didasarkan pada keakuratan setelah melakukan evaluasi menggunakan Persamaan 1.

$$Accuracy = \frac{\sum_{i=1}^{n} N_{ii}}{\sum_{i=1}^{n} \sum_{j=1}^{n} N_{ij}} \quad (1)$$

Perhitungan akan digunakan sesuai dengan penelitian Mehak, dkk. [7]. Penyesuaian dalam presentase seperti ditulis dalam Persamaan 2.

$$Accuracy = \frac{\sum_{i=1}^{n} N_{ii}}{\sum_{i=1}^{n} \sum_{j=1}^{n} N_{ij}} x100\% \quad (2)$$

Hasil dari evaluasi diharapkan dapat menunjukan hasil dari keselurahan proses ekstraksi data yang memiliki hasil penuh atau data dengan atribut lengkap.

4. Hasil dan Diskusi

Fokus dari penelitian ini melakukan proses ekstraksi data dari profil pengguna sebuah situs web *e-commerce*. Ekstraksi data yang dilakukan menggunakan untuk mengumpulkan data dalam empat atribut berdasarkan data publik yang disediakan dan dapat diakses secara umum. Pada setiap hasil data yang telah diambil harapannya memiliki keempat atribut tersebut secara lengkap.

Tahap pertama yaitu menentukan kategori apa saja yang terdapat dalam situs web tersebut kemudian diambil empat kategori utama yang memiliki popularitas paling tinggi di situs web tersebut. Hasil dari pengelompokan tersebut seperti terlihat dalam Tabel 2. Keempat kategori situs web yang akan dijadikan target dalam penelitian ekstraksi data ini yaitu Mobil, Properti, Elektronik & *Gadget* dan Rumah Tangga. Kategori yang sudah ada kemudian akan dimasukkan kedalam daftar yang nanti akan disesuaikan dengan atribut yang yang dilakukan proses ekstraksi data.

Tabel 2. Daftar Kategori Ekstraksi Data

No	Kategori
1	Mobil
2	Properti
3	Elektronik & *Gadget*
4	Rumah Tangga

Terdapat empat atribut dalam penelitian ini berdasarkan pada kebutuhan data ekstraksi yang terdapat dalam masing-masing kategori yang telah ditentukan dalam daftar sebelumnya. Daftar keempat atribut seperti yang terdapat dalam Tabel 3. Diantaranya meliputi nama,

kategori, kontak dan alamat yang masing-masing kami berikan simbol untuk mempermudah dalam proses penelitian, yaitu menjadi nama (N), kategori (K), kontak (H) dan alamat (A). Masing-masing atribut menggambarkan satu data yang akan dilakukan ekstraksi sehingga nantinya setiap hasil ekstraksi akan memiliki empat atribut tersebut.

Tabel 3. Daftar Atribut Ekstraksi Data

No	Atribut
1	Nama (N)
2	Kategori (K)
3	Kontak (H)
4	Alamat (A)

Hasil dari kategori dan atribut yang telah dimasukkan ke daftar sebelumnya kemudian digabungkan keduanya seperti yang ditampilkan dalam Tabel 4. Masing-masing kategori memiliki empat atribut yang akan diekstraksi yaitu kategori mobil dengan atribut (N)(K)(H)(A), kategori properti dengan atribut (N)(K)(H)(A), kategori elektronik dan *gadget* dengan atribut (N)(K)(H)(A) dan kategori rumah tangga dengan atribut (N)(K)(H)(A).

Tabel 4. Daftar Kategori dilenkapi dengan atribut ekstraksi data

No	Mobil	Properti	Elektronik & *Gadget*	Rumah Tangga
1	Nama (N)	Nama (N)	Nama (N)	Nama (N)
2	Kategori (K)	Kategori(K)	Kategori (K)	Kategori (K)
3	Kontak (H)	Kontak (H)	Kontak (H)	Kontak (H)
4	Alamat (A)	Alamat (A)	Alamat (A)	Alamat (A)

Dalam penelitian ini akan menggunakan skenario empat kasus ekstraksi data, setiap kategori yang ada mewakili satu kasus skenario proses ekstraksi data seperti yang ditampilkan dalam Tabel 5. Kategori yang telah disusun dengan atribut akan dijadikan acuan dalam perulangan ekstraksi data. Kasus 1 akan dilakukan proses ekstraksi pada kategori mobil, kasus 2 pada kategori properti, kasus 3 pada kategori elektronik dan *gadget* dan kasus 4 pada kategori rumah tangga.

Tabel 5. Skenario Ekstraksi dalam 4 Kasus

Skenario	Kategori	Atribut 1	Atribut 2	Atribut 3	Atribut 4
Kasus 1	Mobil	Nama (N)	Kategori (K)	Kontak (H)	Alamat (A)
Kasus 2	Properti	Nama (N)	Kategori (K)	Kontak (H)	Alamat (A)
Kasus 3	Elektronik & *Gadget*	Nama (N)	Kategori (K)	Kontak (H)	Alamat (A)
Kasus 4	Rumah Tangga	Nama (N)	Kategori (K)	Kontak (H)	Alamat (A)

Tabel 5 menampilkan daftar atribut yang akan digunakan kedalam tahapan berikutnya, yaitu identifikasi tag yang terdapat di situs web target disesuaikan dengan atribut yang ada. Setiap kategori memiliki atribut yang sama sehingga hanya dibutuhkan satu kali idetifikasi untuk melakukan ekstraksi keempat kasus yang telah disusun.

4.1. Identifikasi Elemen Tag

Proses identifikasi elemen tag dimulai dengan menentukan daftar atribut yang dibutuhkan seperti yang ditampilkan pada Tabel 3. Berikutnya melihat struktur dari situs web yang akan dijadikan target, dari keseluruhan kode yang ada di situs web dilakukan pengecekan pada bagian yang sesuai dengan atribut apakah atribut di situs web tersebut didasarkan pada tag html, selector css atau xpath.

Hasil identifkasi elemen tag kemudian disusun mejadi tag ekstraksi data seperti yang telihat pada Kode 1. Kode tag ekstraksi data menjalankan proses untuk ekstraksi langsung

mengambil keempat atribut yang ada berdasarkan masing-masing kategori yang telah didefinisikan.

Kode 1. Tag Ekstraksi Data

```
post +=  "CODE:";
post += "SET !ERRORIGNORE YES" + "\n";
post += "SET !EXTRACT_TEST_POPUP NO" + "\n";
post += "TAG POS={{loop}} TYPE=A ATTR=HREF:/item/* EXTRACT=HREF" + "\n";
post += "TAB OPEN" + "\n";
post += "TAB T=2" + "\n";
post += "URL GOTO={{!EXTRACT}}" + "\n";
post += "SET !EXTRACT NULL" + "\n";
post += "TAG POS=1 TYPE=DIV ATTR=CLASS:_3oOe9 EXTRACT=TXT" + "\n";
post += "TAG POS=1 TYPE=DIV ATTR=TXT:Tampilkan<SP>nomor" + "\n";
post += "TAG POS=1 TYPE=DIV ATTR=TXT:+628* EXTRACT=TXT" + "\n";
post += "TAG POS=1 TYPE=DIV ATTR=DATA-AUT-ID:itemLocation EXTRACT=TXT" + "\n";
post += "TAG POS=1 TYPE=DIV ATTR=DATA-AUT-ID:breadcrumb EXTRACT=TXT" + "\n";
var load;
load =  "CODE:";
load += "TAG POS=1 TYPE=BUTTON ATTR=TXT:muat<SP>lainnya" + "\n";
load += "WAIT SECONDS=10" + "\n";
```

Dalam penyusunan tag ekstraksi data didasarkan pada elemen tag disesuaikan dengan kebutuhan atribut yang ada, penggunaan `class` dalam `attr` menandakan bahwa dalam proses ekstraksi data untuk salah satu atributnya berdasarkan dari CSS *selector*. Untuk txt sendiri merupakan data yang disajikan dalam tag html dan xpath sendiri digunakan untuk pengambilan berdasarkan tautan data yang ada. Beberapa jenis elemen tag digunakan berdasarkan susunan dari struktur situs web yang dijadikan target.

4.2. Proses Ekstraksi Data

Pendekatan ekstraksi data berdasarkan elemen tag yang telah dilakukan berdasarkan susunan tag ekstraksi data seperti yang terlihat pada Tabel 6. Proses ekstraksi data dilakukan perulangan sebanyak empat kali dengan masing proses berdasarkan kategori masing-masing yang sebelumnya telah di tentukan.

Proses ekstraksi data dijalankan secara otomatis berdasarkan kode tag ekstraksi data yang telah disusun untuk mengambil informasi profil pengguna situs web iklan baris. setiap kasus dijalankan secara bertahap dimulai dari kasus 1 sampai dengan kasus ke empat. Masing-masing kasus ditentukan waktu tertentu dalam ekstraksi data sehingga di hasilkan data yang seharusnya sama. Dalam proses ini setiap pengambilan data dari situs web target didasarkan atribut yang telah disusun, setiap kali dijalankan akan langsung melakukan ekstraksi terhadap empat atribut.

Tabel 6. Hasil Ekstraksi Data

Skenario	Jumlah Data
Kasus 1	4.000
Kasus 2	4.000
Kasus 3	4.000
Kasus 4	4.000
Total	16.000

Total data yang berhasil diekstraksi secara keseluruhan sebanyak 16.000 data dengan empat kali skenario kasus. Pada kasus pertama sampai keempat berhasil mengambil data dengan jumlah yang sama, yaitu 4.000 pada setiap kasusnya. Data yang diambil merupakan data keseluruhan dengan melihat jumlah baris yang berhasil diekstraksi tetapi belum dilakukan proses *filtering* terhadap data yang ada.

4.3. *Filtering* Dataset

Filtering data dilakukan untuk mendapatkan hasil yang sesuai dengan tujuan dari penelitian ini. Hasil dari proses *filtering* data seperti yang ditampilkan pada Tabel 7. Dari total keseluruhan data yang berhasil diekstraksi, dilakukan *filtering* berdasarkan data yang memiliki empat atribut yang lengkap yaitu nama (N), kategori (K), kontak (H) dan alamat (A). Setelah dilakukan, terdapat beberapa perbedaan hasil pada setiap kasus yang dijalankan. Kasus 1 dan kasus 4 memiliki hasil akhir yang sama dengan ekstraksi awal, yaitu masing-masing 4.000 data. Untuk kasus kedua, setelah dilakukan *filtering* terdapat 3.960 data yang lengkap dengan empat atribut. Sedangkan kasus 3 terdapat 3.890 data yang lengkap dengan empat atribut. Keseluruhan data akhir yang memiliki empat atribut secara lengkap terdapat 15.850 data.

Tabel 7. Hasil Ekstraksi Data Lengkap Empat Atribut

Skenario	Jumlah Data
Kasus 1	4.000
Kasus 2	3.960
Kasus 3	3.890
Kasus 4	4.000
Total	15.850

4.4. Evaluasi

Tahapan evaluasi pada penelitian ini dilakukan dengan tujuan untuk mengetahui hasil dari setiap proses ekstraksi data yang dilakukan. Dalam proses ekstraksi data, dihasilkan dua jenis data, yaitu data dari hasil ekstraksi dengan atribut lengkap (1) dan data dari hasil ekstraksi yang memiliki atribut tidak lengkap (0).

Berdasarkan penelitian diperoleh total data 16.000 data dari hasil ekstraksi data secara keseluruhan. Pada ekstraksi data, kasus 1 terdapat 4.000 dengan empat atribut lengkap dan 0 atribut yang tidak lengkap. Pada kasus 2, terdapat 3.960 data hasil ekstraksi dengan empat atribut lengkap dan 40 data dengan atribut tidak lengkap. Sedangkan pada kasus 3, terdapat 3.890 data hasil ekstraksi data dengan empat atribut lengkap dan 110 data hasil ekstraksi yang tidak memiliki atribut lengkap. Hasil terakhir pada kasus 4, terdapat 4.000 data hasil ekstraksi data dengan empat atribut lengkap dan 0 data yang tidak memiliki atribut lengkap.

Hasil evaluasi dilakukan berdasarkan rumus di Persamaan 2 terhadap keseluruhan data hasil proses ektraksi data. Dari perbandingan data hasil ekstraksi dengan atribut lengkap dan data hasil ektraksi dengan atribut tidak lengkap, menghasilkan akurasi proses ekstraksi data sebesar 99.0625%.

5. Kesimpulan dan Saran

Penelitian ini bertujuan untuk melakukan ekstraksi data profil pengguna *website* iklan baris dengan tahapan ekstraksi data berdasarkan elemen tag, yaitu css selector, xpath dan tag html yang disusun kedalam sebuah *extraction tag*. Ekstraksi data yang dilakukan dengan skenario empat kali pengulangan dengan kasus berbeda menghasilkan 16.000 total data. Berdasarkan hasil pengujian yang dilakukan, nilai perhitungan tingkat akurasi mencapai 99.0625%. Untuk hasil yang lebih maksimal, penggabungan dengan metode yang lain dapat dilakukan atau dengan penambahan jumlah data yang lebih banyak.

Referensi

[1] K. Diah and W. Yunanto, "Heuristics miner for e-commerce visitor access pattern representation," *Communication in Sciences and Technology,* vol. 2, no. 1, pp. 1-5, Jun. 2017.

[2] E. Turban and D. King, Eds., *Electronic Commerce - A Managerial And Social Networks Perspective*, 8th ed. Cham, Switzerland: Springer International Publishing, 2015, pp. 7-11.

[3] V. Mitra, H. Suajini and A. B. P. Negara, "Rancang bangun aplikasi web scrapping untuk korpus paralel Indonesia – Inggris dengan metode HTML DOM" *JUSTIN*, vol. 5*,* no. 1*,* pp. 1-6*,* Jan. 2017.

[4] Y. Guo, J. Zhang and X. Chen, "Adaptively extracting structured data from web pages," *2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, 2019, pp. 1524-1525, doi: 10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00221.

[5] H. J. Carey and M. Manic, "HTML web content extraction using paragraph tags," *2016 IEEE 25th International Symposium on Industrial Electronics (ISIE)*, 2016, pp. 1099-1105, doi: 10.1109/ISIE.2016.7745047.

[6] M. A. Wani, S. Jabin, G. Yazdani dan N. Ahmadd, "*Design of imacros-based data crawler and the behavior analysis of Facebook user*" *arXiv preprint*: 1082.09566, 2018.

[7] S. Mehak, R. Zafar, S. Aslam and S. M. Bhatti, "Exploiting Filtering approach with Web Scrapping for Smart Online Shopping: Penny Wise: A wise tool for Online Shopping," *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, 2019, pp. 1-5, doi: 10.1109/ICOMET.2019.8673399.

[8] A. Surahman, A. F. Octaviansyah dam D. Darwis, "Ekstraksi data produk e-*marketplace* sebagai strategi pengolahan segmentasi pasar menggunakan *web crawler*" *Jurnal SISTEMASI*, vol. 9, no. 1*,* pp. 73-81, 2020.

[9] F. Polidoro, R. Giannini, R. L. Conte, S. Mosca and F. Rosseti, "*Web Scrapping Techniques to collect data on consumer electronics and airfares for italian HICP compilation*" *Statistical Journal of the IAOS*, vol. 31, no. 2, pp. 165-176*,* May 2015.

[10] M. Akbar and A. Wibowo, "Ekstraksi tabel HTML bentuk column-row wise ke dalam basis data" *J. Teknol. Informasi dan Ilkom,* vol. 5, no. 6, pp. 653, Dec. 2018.

[11] A. Sasongko, "Integrasi data website student.bsi.ac.id untuk mobile infokampus berbasis Android menggunakan ekstraksi HTML" *J.I.T.K*, vol. 2, no. 2, pp. 146-155, Feb. 2017.

[12] X. Yu and Z. Jin, "Web content information extraction based on DOM Tree and statistical information" *IEEE 17th ICCT*, pp. 1308-1311, Oct. 2017.

[13] E. R. Astanti, A. R. Chrimanto and Y. Lukito," Chrome extension untuk data grabber media sosial Twitter dengan metode XPath selector", *Jurnal Teknologi Informasi*, vol. 19, no. 4, pp. 422-436, Nov. 2020.

[14] F. Handayani and S. Pribadi, "Implementasi algoritma Naïve Bayes Classifier dalam pengklasifikasian teks otomatis pengaduan dan pelaporan masyarakat melalui layanan call center 110," *Jurnal Teknik Elektro*, vol. 7, no. 1, pp. 19-24, Jan. 2015.

[15] V. A. Permadi, "Analisis sentimen menggunakan algoritma Naïve Bayes terhadap review restoran di Singapura," *Jurnal Buana Informatika,* vol. 11, no. 2, Okt. 2020.