

Implementasi *Web Scraping* untuk Pengumpulan Informasi Promo Makanan Menggunakan Klasifikasi *Naïve Bayes*

Dave Julianno¹, Agustinus Noertjahyana², Anita Nathania Purbowo³
Program Studi Informatika, Fakultas Teknologi Industri, Universitas Kristen Petra
Surabaya, Indonesia

Telp. (031) – 2983455, Fax. (031) – 8417658

davejulianno@gmail.com¹, agust@petra.ac.id², anitaforpetra@gmail.com³

ABSTRAK

Pada zaman sekarang, penyampaian informasi promo kepada orang-orang dapat dilakukan dengan mudah. Banyak sekali situs yang menyediakan informasi mengenai promo makanan. Tentu situs tersebut dapat membantu orang-orang untuk mencari promo yang paling menguntungkan. Namun, beberapa situs masih tidak menyediakan fitur pencarian yang lengkap. Selain itu, mungkin saja terdapat promo pada usaha penyedia makanan yang sama, hanya saja dengan harga yang berbeda. Tentu saja orang-orang akan berusaha mencari promo makanan dengan harga yang terbaik, sehingga mereka akan memeriksa situs penyedia promo untuk mencari penawaran terbaik. Hal tersebut memakan waktu yang tidak sedikit, dan juga kurang efisien untuk dilakukan. Proses *Naïve Bayes* juga digunakan untuk menentukan apakah promo tersebut di kategori makanan atau tidak.

Aplikasi yang dibuat melakukan pengumpulan semua informasi promo dari situs-situs penyedia promo. Untuk melakukannya, diperlukan sebuah *server* yang dapat mengambil data promo pada setiap situs penyedia promo. Setelah proses pengambilan, dilakukan proses pengolahan data serta proses klasifikasi apakah promo tersebut di kategori makanan atau tidak. Data yang sudah diproses akan ditampilkan pada aplikasi yang dibuat.

Berdasarkan hasil pengujian yang sudah dilakukan, program yang dibuat berhasil untuk mengumpulkan, mengolah dan menampilkan data tentang sebuah promo dari situs-situs penyedia promo. Proses klasifikasi *Naïve Bayes* juga berhasil diterapkan untuk membedakan promo sesuai kategori, meskipun masih ada kekurangan dalam proses tersebut.

Kata Kunci: *Web Scraping*, *BeautifulSoup*, *Naïve Bayes*, *Promo Makanan*.

ABSTRACT

In this day, the delivery of promo information to people can be done easily. Lots of sites that provide information about food promos. Of course the site can help people find the most profitable promos. However, some sites still do not provide complete search features. In addition, there may be promos on the same food supply business, but with a different prices. Of course people will try to find food promos at the best prices. This takes a lot of time, and also less efficient to do. The Naïve Bayes process is also used to determine wheter the promo is in the food category or not.

The application was made to collect all promo information from promo provider websites. To do this, we need a server that can retrieve promo data on each promo provider site. After the retrieval process, the data processing and classification process is carried

out wheter the promo is in the food category or not. Data that has been processed will be displayed on the application.

Based on the result of test that been done on program, program managed to collecting, processing, and displaying data about a promo from promo provider websites. The Naïve Bayes classification process successfully applied to differentiate promos according to category, although there were still shortcomings in the classification process.

Keywords: *Web Scraping*, *BeautifulSoup*, *Naïve Bayes*, *Food Promo*

1. PENDAHULUAN

Makanan merupakan salah satu dari kebutuhan dasar pada kehidupan sehari-hari. Setiap manusia memerlukan makanan, air, oksigen, dan tempat berlindung untuk bertahan hidup. Apabila salah satu dari kebutuhan tersebut tidak dapat terpenuhi, manusia tidak dapat bertahan hidup [5]. Tingginya kebutuhan makanan dari setiap orang juga menyebabkan naiknya jumlah usaha penyedia makanan. Pertumbuhan usaha penyedia makanan di Indonesia memiliki nilai rata-rata sebesar 1.48% dari tahun 2010 sampai 2013. Nilai tersebut berada di atas nilai rata-rata pertumbuhan industri kreatif sebesar 0.98% dan pertumbuhan nasional sebesar 1.05% [4]. Tingginya pertumbuhan usaha penyedia makanan ini menyebabkan persaingan antar usaha semakin tinggi pula. Banyak cara yang dilakukan oleh pemilik usaha untuk menarik perhatian orang-orang, salah satunya adalah pengadaan promo.

Promo diberikan oleh pemilik sebuah usaha kepada konsumen yang tertarik dengan produk yang disediakan. Dengan memberikan promo yang menguntungkan, orang-orang akan tertarik dengan produk yang disediakan, menyebabkan banyak orang akan mengenal produk yang disediakan. Promo pada makanan pada umumnya akan memberikan harga yang lebih murah dari harga asli makanan tersebut, dengan syarat dan ketentuan yang sudah ditetapkan. Seseorang dapat membeli makanan dengan harga yang lebih murah daripada harga asli makanan tersebut.

Pada zaman sekarang, penyampaian informasi promo kepada orang-orang dapat dilakukan dengan mudah. Banyak sekali situs yang menyediakan informasi mengenai promo makanan. Tentu situs tersebut dapat membantu orang-orang untuk mencari promo yang paling menguntungkan. Namun, beberapa situs masih tidak menyediakan fitur pencarian yang lengkap. Selain itu, mungkin saja terdapat promo pada usaha penyedia makanan yang sama, hanya saja dengan harga yang berbeda. Tentu saja orang-orang akan berusaha mencari promo makanan dengan harga yang terbaik, sehingga mereka akan memeriksa situs penyedia promo untuk

mencari penawaran terbaik. Hal tersebut memakan waktu yang tidak sedikit, dan juga kurang efisien untuk dilakukan. Oleh karena itu, diperlukan aplikasi yang memudahkan pengumpulan informasi promo makanan.

Promo yang disediakan pada situs penyedia promo tidak hanya di kategori makanan saja. Banyak sekali promo di kategori yang beragam disediakan oleh situs. Untuk memilih dan mengklasifikasi promo yang termasuk kategori makanan, digunakan metode *machine learning*. Proses *machine learning* digunakan untuk menentukan klasifikasi dari sebuah data dengan melakukan *training* dan *testing* [6]. Informasi mengenai setiap promo yang ada akan diklasifikasikan dengan menggunakan metode *Naïve Bayes*. Metode *Naïve Bayes* menggunakan konsep probabilitas, semakin tinggi probabilitas data terhadap klasifikasi, maka kemungkinan data tersebut terklasifikasikan di kelas yang sama.

Web scraping adalah salah satu cara untuk mengambil informasi mengenai promo pada sebuah situs. *Web scraping* akan mengambil data yang ada pada *World Wide Web (WWW)* kemudian akan menyimpan data tersebut ke sistem *file*, seperti *database* [2]. Banyak metode yang bisa digunakan untuk melakukan *web scraping*. Metode *human manual copy-and-paste* merupakan salah satu metode yang paling sering digunakan untuk memperoleh sebuah data dari situs. Data yang ingin diambil akan di-*copy* secara manual oleh manusia, kemudian akan dipindahkan sesuai struktur data yang sesuai. Sayangnya, metode ini tidak dapat memproses data dengan skala yang besar, dan juga membutuhkan waktu yang relatif lama.

2. TINJAUAN STUDI

Web scraping adalah salah satu cara untuk mengambil informasi mengenai promo pada sebuah situs penyedia promo. *Web scraping* akan mengambil data yang ada pada *world wide web (WWW)* kemudian akan menyimpan data tersebut ke sistem *file*, seperti *database* [2]. Seringnya terjadi kesalahan terhadap *web scraping* dengan *web crawling*. *Web crawling* tidak perlu mendefinisikan target yang diinginkan, sehingga *web crawling* akan memproses semua data yang ada tanpa meminta informasi yang spesifik. Sedangkan untuk *web scraping* melakukan proses data dari sumber yang spesifik. Salah satu cara untuk melakukan *web scraping* adalah dengan menggunakan BeautifulSoup.

BeautifulSoup adalah sebuah Python *database* yang berdasarkan penemuan mesin analisa HTML dan XML, digunakan untuk melakukan ekstraksi, analisa, dan mengubah informasi di dalam pohon DOM dari sebuah situs [1]. BeautifulSoup akan menganalisa semua dokumen yang diberikan, termasuk semua jenis dokumen HTML dan XML. Data yang sudah berhasil didapat kemudian diklasifikasikan menggunakan *Naïve Bayes*.

Algoritma *Naïve Bayes* merupakan sebuah metode klasifikasi yang berjalan dengan konsep probabilitas dan statistic yang dikemukakan oleh ilmuwan Inggris Thomas Bayes. *Naïve Bayes* adalah sebuah algoritma untuk melakukan klasifikasi berdasarkan dalil Bayes dengan asumsi yang kuat dan independen [3]. Algoritma ini memprediksi peluang di masa depan berdasarkan nilai nilai yang terjadi di pengalaman di masa sebelumnya. Algoritma *Naïve Bayes* ini dikenal karena memiliki asumsi yang sangat kuat atau naif terhadap independensi dari masing-masing kondisi atau kejadian.

3. DESAIN SISTEM

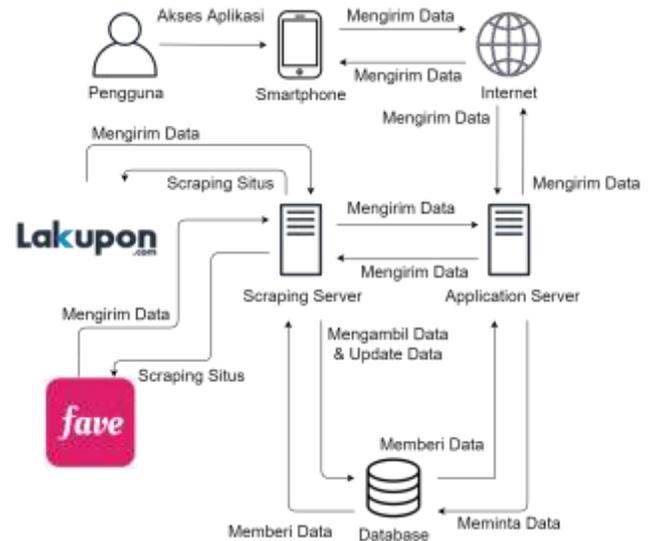
Proses dalam pembuatan sistem meliputi proses *scraping* ke situs-situs penyedia promo, memproses data, pembuatan model untuk

Naïve Bayes, *training* dan *testing* data dari model yang telah dibuat, penggunaan model untuk klasifikasi kategori promo, dan desain *user interface*.

Terdapat *back end* dan *front end* dalam pembuatan sistem, *back end* bertugas untuk memproses data promo yang sudah di-*scraping*. Sedangkan *front end* bertugas untuk menampilkan hasil dari data yang sudah diolah.

3.1 Arsitektur Sistem

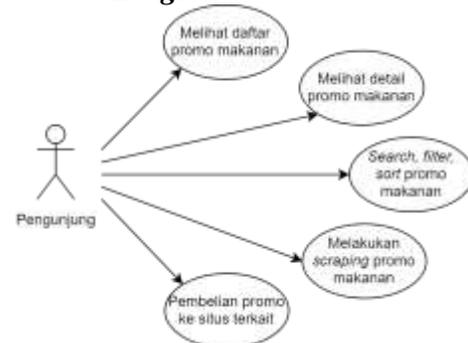
Untuk mengetahui bagaimana sebuah sistem bekerja, dibutuhkan sebuah desain struktur dari sistem tersebut. Hal ini bertujuan agar pengguna bisa memahami proses yang terjadi.



Gambar 1. Arsitektur sistem

Pada gambar 1, aplikasi akan mengambil data promo yang ada di *database* dan menampilkannya kepada pengguna. Data dari *database* merupakan data yang sudah didapatkan melalui proses *scraping* sebelumnya. Apabila pengguna ingin mendapatkan promo terbaru dari situs penyedia promo, sudah tersedia tombol untuk melakukan *scraping* data promo baru. Pada saat tombol ditekan, aplikasi akan memanggil *server* untuk melakukan *scraping* ke situs penyedia promo. Apabila proses *scraping* sudah selesai, maka data hasil *scraping* akan diproses ke dalam *database*. Setelah itu, data dari *database* akan diproses kembali untuk klasifikasi jika data tersebut masih belum diklasifikasikan. Setelah selesai melakukan klasifikasi, maka dilakukan *update* data yang baru ke *database*. Data tersebut kemudian diakses oleh aplikasi untuk ditampilkan kepada pengguna.

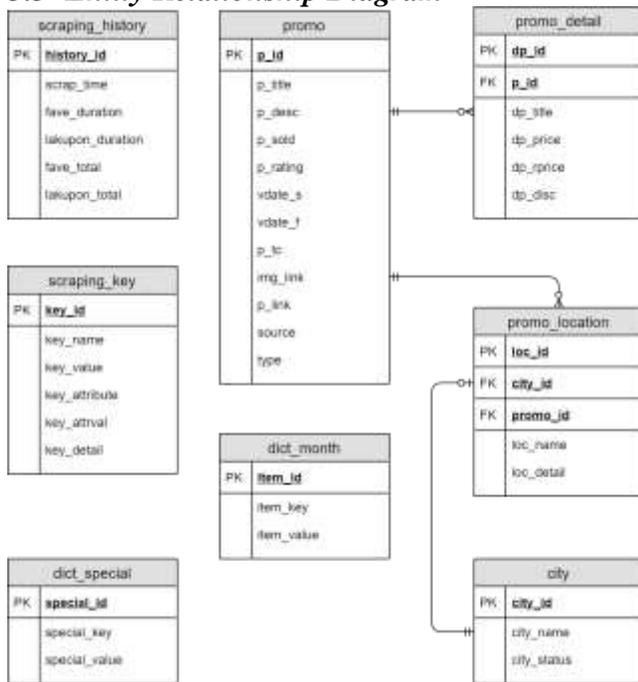
3.2 Use Case Diagram



Gambar 2. Use case diagram sistem

Pada gambar 2, terdapat aktor, pengunjung. Pengunjung bisa melihat daftar promo makanan yang tersedia dari beberapa situs penyedia promo. Selain itu, pengunjung juga bisa menjalan proses *scraping* dengan menekan tombol yang ada di aplikasi, sehingga data promo yang baru dibuat bisa ditampilkan di aplikasi. Pengunjung bisa mencari promo makanan yang diinginkan dengan memasukkan nama outlet. Pengunjung bisa mengurutkan daftar promo berdasarkan harga, jumlah terjualnya promo, dan sesuai urutan *alphabet*. Pengunjung juga bisa melakukan penyaringan atau *filter* berdasarkan harga promo, lokasi promo, dan juga periode berlaku promo.

3.3 Entity Relationship Diagram



Gambar 3. Entity relationship diagram sistem.

Data yang diperlukan dalam menjalankan sistem ini disimpan dan diatur di dalam *database*. Agar data disimpan dengan efisien dan optimal, maka diperlukan susunan *database* yang bagus dan jelas, sehingga tidak terjadi kesalahan atau kehilangan data akibat struktur *database*.

Pada gambar 3, untuk setiap promo yang akan disimpan, data dari promo akan dimasukkan ke dalam tabel *promo*. Data yang disimpan di tabel ini antara lain adalah judul promo, deskripsi promo, jumlah promo terjual, rating promo, masa berlaku promo, syarat ketentuan promo, sumber promo, tipe promo, dan juga *link* menuju ke situs asli promo. Setiap data promo akan memiliki setidaknya satu data detail promo. Data untuk detail promo berisikan detail promo yang ditawarkan, apakah promo hanya memberikan satu jenis penawaran saja kepada pengguna. Detail promo berisikan data berupa judul detail promo, harga dari detail promo, harga asli dari detail promo, besar diskon promo. Selain memiliki detail promo, data promo juga memiliki setidaknya satu data lokasi tempat promo tersebut bisa digunakan. Data yang disimpan untuk tabel lokasi antara lain adalah nama lokasi, serta detail dari lokasi.

3.4 Scraping Data

Situs yang telah ditentukan akan diambil data promo yang diperlukan menggunakan proses *scraping*. Setiap situs memiliki karakteristik dan struktur data yang berbeda satu dengan yang lain.

Oleh karena itu diperlukan analisa struktur data di setiap situs. Pertama tama perlu dicari data apa saja yang diperlukan. Setiap data yang sekiranya diperlukan, kemudian dicari lokasi tempat tersimpannya data tersebut. Apabila setiap data yang diperlukan sudah teridentifikasi letak pada halaman situs tersebut, maka dilakukan proses *scraping*. Untuk setiap promo yang berhasil diambil, perlu dilakukan pemeriksaan. Apabila data yang baru diambil sudah ada di dalam *database* promo, maka data baru tersebut tidak perlu diproses lebih jauh lagi. Namun, apabila data yang baru diambil belum ada di dalam *database* promo, maka data baru tersebut akan diproses lebih lanjut, kemudian data tersebut ditambahkan ke dalam *database*.

```

    [{"average_rating": "4.90", "company_id": "328", "company_name": "Dairy Queen", "description": "Dairy Queen kini hadir di Fave dengan penawaran Buy 1 Medium Blizzard Oreo Flavor Get 1 Medium (Any Flavor). Dairy Queen menawarkan ice cream super-premium dengan berbagai macam pilihan rasa seperti Cappuccino Oreo Blizzard, Oreo Green Tea Blizzard, Strawberry Oreo Blizzard, dan masih banyak lagi. DIBuat dari bahan baku berkualitas, dan diproduksi setiap harinya dapat menjadikan Dairy Queen menjadi hidangan penutup yang cocok untuk anda nikmati bersama keluarga, dan kawan-kawan terdekat.", "discount": "50", "discounted_price": "Rp58.000", "discounted_price_cents": "5800000", "distance": "665.5 km", "end_date": "2020-06-30T23:59:59.999+08:00", "featured_thumbnail_image": "https://image-assets.access.myfave.gdn/attachments/c10124cc9536526754c2401469ec"}
    
```

Gambar 4. Struktur HTML situs promo

Analisa struktur dari sebuah situs penyedia promo adalah seperti pada gambar 4. Data yang diperlukan disimpan dengan format data berupa JSON. Oleh karena itu, data tersebut diolah terlebih dahulu sehingga data yang diperlukan bisa diambil sesuai kebutuhan.

```

    outlets: [{"activities": [], "address": "Jl. Penuda no. 118 Semarang - Jawa Tengah 50132", "address_latitude": "-6.979298", "address_longitude": "110.415967", "city_id": "5", "company": null, "company_id": "5783", "deleted_at": null, "email": "Smggk26@map.co.id", "fave_payments": [], "favorited_count": "0", "featured": false, "has_fave_payment": false, "id": "24726", "name": "Paragon Mall Semarang, Lt 2", "offers_count": "7", "status": "active", "telephone": "", "town_name": null, ("activities": [], "address": "Jl. Puri Agung, Puri Indah Jakarta 11610", "address_latitude": "-6.188302", "address_longitude": "106.734185", "city_id": "5", "company": null, "company_id": "5783", "deleted_at": null, "email": "Jakgk25@map.co.id", "fave_payments": [], "favorited_count": "0", "featured": false, "has_fave_payment": false,
    
```

Gambar 5. Struktur HTML lokasi promo

```

    <div class="title-detail">SYARAT DAN KETENTUAN</div>
    <div class="content-detail" style="margin-left: 20px">
    <li>Pembayaran ditunggu Max. 2 x 24 Jam, apabila tidak ada pembayaran dalam kur maka order akan otomatis terhapus oleh siste. Silakan lakukan pemesanan kembali.</li>
    <li>Harga belum termasuk Ongkos Kirim, Ongkos kirim bervariasi sesuai dengan lo</li>
    <li>Apabila memilih pengambilan di kantor Lakupon customer wajib membawa kupon dicetak dan 1 (satu) buah fotokopi identitas diri untuk proses verifikasi pengambil</li>
    <li>Apabila memilih opsi delivery, barang akan dikirim sesuai tanggal awal peng pada ketentuan deal (3 hari apabila ada logo 3 Day Delivery)</li>
    <li>Produk tidak dapat diambil jika sudah melewati periode pengambilan yang tel customer tidak dapat mengklaim apapun ke pihak Lakupon.</li>
    
```

Gambar 6. Struktur HTML syarat ketentuan promo

Penggunaan BeautifulSoup diterapkan untuk melakukan pengambilan data dari situs promo. BeautifulSoup akan menyimpan *link* dari setiap situs, kemudian digunakan untuk mengakses halaman situs. BeautifulSoup juga digunakan untuk mengambil data yang berada di dalam *tag*, dengan mengenali data yang berada di dalam *tag* tersebut.

Penggunaan *regular expression* digunakan untuk mencari data promo dengan pola kalimat yang sesuai. *Regular expression* akan mencari kata atau kalimat yang sesuai dengan pola yang sudah dibuat. Penggunaan *regular expression* diterapkan untuk mengambil tanggal berlaku promo, karena tanggal berlaku memiliki format yang selalu sama dan tidak pernah berganti.

3.5 Naïve Bayes

Setelah proses *scraping* selesai dijalankan di kedua situs yang ada, maka proses klasifikasi akan dijalankan. Sebelumnya, sudah ditentukan bahwa data promo yang akan diklasifikasikan adalah data promo yang berasal dari situs Lakupon. Hal itu disebabkan karena data promo yang diambil dari situs Lakupon masih tercampur dengan data promo yang lain. Oleh karena itu, ketika proses *scraping* dilakukan, maka data promo yang diambil dari situs Fave sudah berada di kategori makanan, sedangkan data promo yang diambil dari situs Lakupon belum ditentukan kategorinya. Algoritma Naïve Bayes akan diterapkan pada data promo yang diambil dari situs Lakupon.

Pertama diperlukan pembuatan model yang didapatkan dari data promo dengan kategori makanan dan data promo dengan kategori bukan makanan. Data ini didapatkan dari situs Fave, sehingga model Naïve Bayes ini dirancang berdasarkan data dari situs Fave. Setelah mendapatkan model, maka setiap data promo dari situs Lakupon diterapkan algoritma Naïve Bayes dengan model yang sudah dibuat. Perhitungan probabilitas setiap data dilakukan dengan menghitung jumlah setiap kata yang ada pada deskripsi promo terhadap semua kata yang ada pada model. Apabila probabilitas promo tersebut di kategori makanan lebih besar daripada probabilitas di kategori bukan makanan, maka promo tersebut bisa dikategorikan ke kategori makanan, begitu pula sebaliknya.

4. PENGUJIAN SISTEM

4.1 Kesesuaian Data Promo



Gambar 7. Contoh tampilan aplikasi

Pada gambar 7 merupakan tampilan dari aplikasi yang menampilkan detail dari promo makanan yang bersangkutan. Data dari situs Fave berhasil diambil dengan aman, tanpa ada kesalahan atau hilang.



Gambar 8. Tampilan detail promo pada aplikasi

Dari gambar 8 dan didapatkan bahwa data yang ditampilkan di situs Lakupon berhasil diambil dengan baik dan lancar. Data yang ada di situs Lakupon sama dengan data yang ditampilkan pada aplikasi.

4.2 Durasi Scraping

Tabel 1. Daftar durasi detail *scraping*

Keterangan	Waktu Fave	Waktu Lakupon	Total Fave	Total Lakupon
Detail <i>scraping</i> ketika belum ada data yang tersimpan	592.84	219.29	220	148
Detail <i>scraping</i> ketika sudah ada data yang tersimpan	248.44	201.55	75	41

Dari tabel 1, terdapat 2 data yang tersimpan. Pada data pertama tersimpan detail proses *scraping* ketika belum ada data yang tersimpan di dalam *database*. Pada data kedua tersimpan detail proses *scraping* ketika sudah ada beberapa data yang tersimpan di

- [6] Willcock, S., Martínez-López, J. 2018. Machine learning for ecosystem services. *Ecosystem Services*, 33, 165-174.