MIND Journal | ISSN (p): 2528-0015| ISSN (e): **2528-0902** | Vol. IV | No. 1 | Halaman 01-14
DOI : https://doi.org/10.26760/mindjournal Juni 2019

Rancang Bangun Mesin *Crawler* di Instagram dan Pinterest untuk Kebutuhan Data pada Riset Visual

Halimatus Sa'dyah, Widi Sarinastiti, Reza R Ramadhan
Politeknik Elektronika Negeri Surabaya
Email : [1]halimatus@pens.ac.id

ABSTRAK

Media sosial memberikan bonus berupa data yang dapat dikelola menjadi informasi yang bermanfaat. Dalam penelitian ini, kami mengembangkan mesin crawler untuk media sosial Instagram dan Pinterest. Mesin crawler ini digunakan sebagai infrastruktur pendukung untuk mengambil data di media sosial. Data yang dihasilkan oleh mesin crawler selanjutnya digunakan sebagai bahan riset visual untuk merancang kemasan produk bagi konsumen menengah ke atas. Hasil uji coba menunjukkan bahwa penggunaan Apache MesOS dapat mempercepat proses crawling dari 30 jam menjadi 1 jam. Dalam hal seleksi data, pada Pinterest, mesin crawler ini dapat mencapai akurasi hingga 100%. Sementara itu, pada Instagram, nilai Presisi tidak stabil dan berada pada rentang 34.8% hingga 90.0%. Sedangkan nilai recall dan akurasinya konsisten di bawah 70%. Hal ini menunjukkan bahwa arsitekturr mesin crawler sudah sesuai untuk menyelesaikan permasalahan. Namun, perbaikan masih dibutuhkan dari sisi algoritma seleksi agar nilai Presisi, Recall dan Akurasi pada Instagram dapat ditingkatkan lagi.

***Kata kunci**: Media Sosial, Riset Visual, Mesin Crawler, Infrastruktur Big Data*

ABSTRACT

Social media gives us a huge number of data to be analyzed and sends us useful knowledge. In this paper, we develop a crawler machine for Instagram and Pinterest as an infrastructure of social media based visual research. We conduct the visual research to design product package for consumers from the middle class and upper class. The crawler machine is developed using Apache Kafka, Apache Spark, and Apache MesOS The evaluation result shows us that Apache MesOS can speed up data processing from 30 hours to one hour. In term of data selection, this machine can achieve 100% accuracy on Pinterest. Meanwhile, on Instagram, the precision is unstable between 34.8% to 90.0%. On the other hand, the recall and the accuracy on Instagram are consistently below 70%. Based on the evaluation result, we conclude that the machine performs well in term of time efficiency. However, we need to improve the selection algorithm so that the precision, the recall and the accuracy on Instagram can be increased.

Keywords: *Social Media, Visual Research, Crawler Machine, Big Data Infrastructure*

1. PENDAHULUAN

1.1 Latar Belakang Masalah

Persaingan popularitas merek (*brand*) menjadi salah satu penentu apakah sebuah produk dapat bersaing di pasar. Untuk dapat berkompetisi dengan kompetitor, produsen selalu mencoba memberikan hal baru pada produk mereka. Dalam persaingan pasar, konsumen bukan hanya mempertimbangkan aspek kualitas produk, melainkan juga kesan pertama pada kemasannya. Merek yang baik biasanya memahami fungsi kemasan dalam menyampaikan pesan yang diusung oleh produk. Seperti yang disebutkan oleh Rundh **(Rundh, 2009)** bahwa kemasan produk berhubungan dengan variabel lain dalam pemasaran. Kemasan dan desain kemasan menjadi salah satu faktor penentu sebuah produk diminati atau tidak oleh pasar.

Pada perkembangannya, kemasan produk memberikan daya tarik pada konsumen agar membeli sebuah produk. Pada abad ke 21, kecenderungan konsumen untuk membeli barang tidak hanya berdasarkan kebutuhan **(Pine & Gilmore, 2011)**. Perilaku konsumen telah mengalami pergeseran dimana mereka membeli produk dengan mempertimbangkan aspek immaterial seperti emosi, hiburan dan fantasi. Dalam teori marketing tradisional, kemasan merupakan salah satu variable keputusan penting dalam menentukan pembelian **(Pickton & Broderick, 2005) (Kotler & Keller, 2011)**. Dalam kemasan, hal terbesar yang menjadi pendorong konsumen untuk dapat tertarik kepada sebuah produk salah satunya adalah pada desain atau ilustrasi yang menerangkan produk yang dijual. Dalam beberapa penelitian sebelumnya, disebutkan bahwa visual influence terbukti mempengaruhi keputusan pembelian suatu barang **(Clement, Kristensen, & Gronhaug, 2013).**

Dalam disiplin ilmu Desain Komunikasi Visual, desain kemasan dikerjakan melalui metode riset visual. Riset visual memiliki tiga aspek yaitu imaji, pembuat dan pemirsa. Aspek imaji merujuk pada karya visual yang sedang diteliti yaitu kemasan yang sedang dirancang. Aspek pembuat merujuk pada pembuat karya visual yaitu desainer kemasan. Sedangkan aspek pemirsa merujuk pada penikmat karya visual atau konsumen produk **(Sowardikoen, 2013)**.

Sebuah karya visual tidak terlepas dari persepsi pemirsa terhadap karya tersebut. Oleh karena itu, untuk menghasilkan kemasan yang sesuai dengan selera dan kebutuhan konsumen (sebagai pemirsa), desainer kemasan harus melakukan penggalian informasi tentang konsumen. Pada umumnya, penggalian informasi tentang selera konsumen dilakukan melalui *Focus Group Discussion,* wawancara, atau pembagian kuesioner. Namun, perkembangan teknologi memungkinkan para desainer untuk menggunakan teknik baru dalam menggali informasi.

Tren saat ini menunjukkan banyak penduduk Indonesia yang menggunakan media sosial. Hal memberikan kita bonus berupa data dalam jumlah besar. Kementrian Komunikasi dan Informasi (Kominfo) merilis laporan bahwa jumlah pengguna internet (warganet) di Indonesia saat ini mencapai 63 juta jiwa dimana 95% di antaranya menggunakan internet untuk mengakses media sosial **(Kominfo, 2017)**. Jika setiap pengguna media sosial mengunggah status satu kali per hari, maka setiap harinya kita akan mendapatkan 59 juta rekaman baru. Faktanya, hampir setiap warganet di Indonesia memiliki lebih dari satu akun media sosial. Artinya, data baru yang diperoleh setiap harinya diperkirakan lebih dari 59 juta rekaman. Kondisi

tersebut memungkinkan penggalian informasi tentang pemirsa dilakukan melalui media sosial.

Saat ini, terdapat dua media sosial yang dapat dieksplorasi untuk kebutuhan riset visual yaitu Instagram dan Pinterest. Instagram merupakan media sosial yang digunakan untuk berbagi foto. Pada perkembangannya, Instagram juga digunakan sebagai media promosi produk. Implikasinya, kita dapat mengamati interaksi antar pengguna yang membahas produk-produk tertentu. Dari sini, kita dapat mengamati perilaku dan persepsi mereka terhadap sebuah produk.

Adapun Pinterest merupakan media sosial yang banyak digunakan untuk mengunggah karya visual. Unggahan di Pinterest tidak langsung berkaitan dengan konsumen. Namun, dengan banyaknya desainer grafis yang mengunggah karyanya di sana, kita dapat mengambil informasi tentang tren desain grafis pada tahun tertentu melalui Pinterest.

Dalam penelitian ini, kami membangun mesin *crawler* pada media sosial Instagram dan Pinterest. Pembuatan mesin *crawler* ini ditujukan untuk memenuhi kebutuhan desainer kemasan dalam menggali informasi tentang konsumen. Terdapat dua tantangan yang harus ditempuh jika kita ingin mengembangkan perangkat lunak untuk menggali informasi pada media sosial **(Zafarani, Abbasi, & Huan, 2014**). Tantangan pertama adalah pengembangan teknik pemilihan sampel yang tepat. Data di media sosial terus bertambah dalam hitungan detik. Variasi dan deviasinya hampir tidak terukur sehingga kita akan kesulitan untuk menemukan titik sampel yang tepat. Dalam penelitian ini, penyelesaian untuk permasalahan tersebut dimulai dengan mengumpulkan data pengguna media sosial sesuai dengan segmen konsumen yang dituju. Adapun pengumpulan data tentang pengguna media sosial dilakukan menggunakan algoritma *Rule Based Classifier.*

Tantangan kedua adalah pengembangan infrastruktur yang memadai untuk mengelola data. Proses pengumpulan data pengguna media sosial dilakukan melalui *Application Programming Interface* (API). API menyediakan perintah untuk mengambil data yang dibutuhkan oleh pengguna. Namun, untuk mengumpulkan data yang sesuai kriteria, kita membutuhkan teknik komputasi yang kompleks. Selain itu, data di media sosial juga bertambah dengan cepat dan berukuran besar. Oleh karena itu, kita tidak dapat memproses data media sosial dalam komputer *single node*. Pada penelitian ini, kami menyelesaikan permasalahan tersebut dengan komputasi terdistribusi yang dibangun dengan Apache MesOS.

Makalah ini dibagi menjadi lima bagian. Bagian pertama menjelaskan tentang pendahuluan. Bagian kedua landasan teori yang mendasari penelitian ini. Bagian ketiga menjelaskan tentang penggunaan algoritma *Rule Based Classifier* serta arsitektur mesin *crawler*. Bagian keempat menjelaskan tentang uji coba dan evaluasi. Bagian kelima menjelaskan tentang kesimpulan.

1.2 Rumusan Masalah

Terdapat dua permasalahan yang harus diselesaikan dalam penelitian ini. Permasalahan pertama adalah algoritma seperti apa yang harus digunakan agar mesin *crawler* dapat mengumpulkan data dari segmen konsumen yang dituju. Permasalahan kedua adalah bagaimana merancang arsitektur mesin *crawler* untuk menggali data dari Instagram dan Pinterest dengan mempertimbangkan ukuran dan kecepatan penambahan data.

1.3 Tujuan Penelitian

Adapun tujuan penelitian yang akan dicapai adalah dihasilkannya mesin *crawler* pada Instagram dan Pinterest yang dapat mengunduh data dengan cepat sesuai dengan kriteria data yang ditentukan.

1.4 Ruang Lingkup Penelitian

Ruang lingkup penelitian ini meliputi penentuan algoritma klasifikasi untuk mendapatkan data yang dibutuhkan serta perancangan arsitektur mesin crawler yang dapat memfasilitasi pengambilan data dalam jumlah banyak dan dalam waktu yang cepat. Adapun data rujukan untuk produk yang biasa digunakan oleh konsumen menengah ke atas telah disediakan sebelumnya oleh desainer kemasan.

2. LANDASAN TEORI

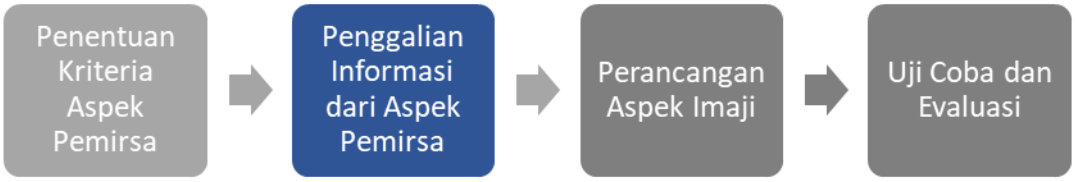
Dalam penelitian ini akan dibuat mesin *crawler* untuk memenuhi kebutuhan riset visual dalam menggali selera konsumen. Oleh karena itu, dibutuhkan landasan teori untuk memahami studi kasus yang dikerjakan serta mengetahui infrastruktur yang akan digunakan. Landasan teori di Bab 2 ini terdiri dari empat bagian yaitu Riset Visual, *Application Programming Interface* (API), infrastruktur *Big Data* dan algoritma *Rule Based Classifier.*

2.1 Riset Visual dan Elemen Kemasan

Kemasan produk memiliki enam fungsi. Fungsi tersebut antara lain fungsi proteksi, fungsi pengelompokan, penempatan dan penyimpanan, keamanan, fungsi informasi, fungsi kemudahan fisik dan fungsi marketing **(Kusrianto, 2007).** Sementara itu, dalam riset visual, kita mengenal tiga aspek yaitu aspek imaji, aspek pemirsa dan aspek pembuat. Kemasan yang baik harus memiliki aspek imaji yang memenuhi selera aspek pemirsa serta memenuhi 6 fungsi yang telah disebutkan

Untuk mendukung keenam fungsi yang telah disebutkan Adi Kusrianto, terdapat lima elemen yang harus dipertimbangkan dalam perancangan kemasan **(Vyas, 2015).** Kelima elemen tersebut adalah nama brand, bahan kemasan, bentuk kemasan, informasi produk dan elemen grafis (kombinasi warna, tata letak dan ukuran). Aspek Imaji dari kemasan dibentuk oleh kombinasi dari lima elemen kemasan ini. Oleh karena itu dibutuhkan informasi yang memadai tentang bagaimana kombinasi elemen kemasan yang memenuhi selera aspek pemirsa. Dalam riset visual ini, aspek pemirsa adalah kalangan menengah ke atas.

Sebelum membuat mesin *crawler,* penting bagi kita untuk mendefinisikan bagaimana ciri konsumen menengah atas di media sosial. Berdasarkan ciri tersebut, mesin *crawler* akan mengambil data di media sosial. Dengan data tersebut, kita dapat memperhatikan perilaku segmen konsumen yang dituju. Pada umumnya, segmentasi konsumen dibagi berdasarkan jumlah pengeluaran mereka setiap bulan. Namun dalam riset menggunakan media sosial, kita tidak mungkin mendapatkan data tersebut. Oleh karena itu, dalam penelitian ini, pembagian segmentasi konsumen dilakukan berdasarkan produk yang mereka konsumsi dan dibagikan melalui media sosial.



Gambar 1. Metodologi Riset Visual

Dalam penelitian ini, data yang dihasilkan oleh mesin *crawler* akan digunakan sebagai bahan riset visual untuk mempertimbangkan kombinasi elemen kemasan yang baik dan dapat diterima oleh konsumen. Secara umum, metodologi riset visual dapat dilihat pada Gambar 1. Pada gambar tersebut, terdapat empat langkah yang harus ditempuh mulai dari penentuan kriteria aspek pemirsa hingga uji coba dan evaluasi. Penelitian yang ada pada makalah ini berfokus pada langkah kedua yaitu penggalian informasi dari aspek pemirsa.

Pada umumnya, penggalian informasi tentang perilaku pemirsa digali melalui kuesioner, *Focus Group Discussion,* atau wawancara. Namun dalam penelitian ini, penggalian informasi dilakukan dengan pendekatan lain. Desainer kemasan akan terlebih dahulu mengumpulkan data berupa kumpulan *username* dari pengguna Instagram yang termasuk segmen konsumen menengah ke atas. Selain itu, desainer kemasan juga mengumpulkan tren desain untuk makanan melalui Pinterest. Penelitian ini dilakukan untuk mendukung otomatisasi pengumpulan data tersebut.

2.2 API pada Instagram dan Pinterest

Pada penelitian ini, kita menggunakan Instagram dan Pinterest sebagai sumber data. Instagram merupakan media sosial yang memfasilitasi penggunanya untuk menunggah foto. Instagram memiliki dua fitur utama yaitu unggah foto di halaman utama dan unggah foto di Instagram Story. Saat ini, Instagram banyak digunakan untuk keperluan jual-beli, *branding* dan periklanan.

Adapun Pinterest merupakan media sosial yang juga memfasilitasi penggunanya untuk mengunggah foto. Namun segmen pengguna Pinterest berbeda dengan Instagram. Pinterest banyak digunakan untuk menyimpan portofolio foto ataupun desain grafis. Oleh karena itu, Pinterest sering digunakan sebagai rujukan untuk mencari inspirasi desain oleh para desainer grafis. Pinterest menyebut setiap gambar yang diunggah sebagai Pin. Pin ini dapat dikumpulkan ke dalam suatu album yang disebut dengan Board. Pada umumnya, pengguna Pinterest mengumpulkan Pin dengan topik-topik yang sama ke dalam suatu Board.

Untuk mengumpulkan data dari Pinterest dan Instagram, kita menggunakan API. API *(Application Programming Interface)* merupakan sebuah teknologi yang memungkinkan dua buah aplikasi atau lebih untuk bertukar data tanpa memberikan akses bagi pengguna ke dalam tempat penyimpanan data secara langsung. Instagram dan Pinterest merupakan aplikasi media sosial yang menyediakan API agar kita dapat mengakses data dengan mudah.

Untuk menggunakan API pada masing-masing media sosial, kita dapat menggunakan format *request* tertentu. Format *request* data pada API di media sosial Instagram dapat diakses pada halaman web https://www.instagram.com/developer/. Sedangkan format request data pada API di media sosial Pinterest dapat diakses pada halaman web https://developers.pinterest.com/

2.3 Infrastruktur Big Data

Infrastruktur teknologi adalah hal paling mendasar yang harus terkondisikan dengan baik pada penelitian di bidang *Big Data*. Pada bidang ini, pekerjaan yang harus dilakukan pada umumnya berhubungan dengan penyimpanan, *streaming* dan analisa data. Dalam beberapa dekade, pengelolaan data bergantung pada basis data relasional. Namun, pada kasus *Big Data*, data yang dikelola memiliki kecepatan, volume dan variasi yang sangat besar. Oleh karena itu, basis data relasional tidak lagi mampu untuk mendukung pemrosesan data yang sudah memenuhi kriteria *Big Data.*

Terdapat beberapa infrastruktur yang harus disiapkan saat kita ingin mengelola *Big Data*. Pertama adalah infrastruktur yang menangani aliran data pada *Big Data* yang mengalir secara cepat dan *real time*. Kedua adalah infrastruktur yang menangani penyimpanan data sehingga mudah dikelola. Data pada *Big Data* memiliki banyak variasi. Ada yang terstruktur dan ada yang tidak. Selain itu, data yang disimpan juga berukuran besar. Oleh karena itu, dibutuhkan infrastruktur yang mampu mengatasi hal ini. Ketiga adalah infrastruktur untuk melakukan analisa data. Dan keempat adalah infrastruktur yang menangani pemrosesan data secara terdistribusi. Data pada *Big Data* berukuran besar sehingga membutuhkan waktu yang lama untuk diproses. Adanya pemrosesan terdistribusi memungkinkan kita untuk mempersingkat waktu pengolahan.

Pada penelitian ini, infrastruktur untuk menangani aliran data dibangun dengan menggunakan Apache Kafka*.* Apache Kafka merupakan *platform* untuk *streaming* dengan latensi rendah dan *throughput* tinggi. Apache Kafka biasa digunakan untuk mengumpulkan aliran data secara *realtime* serta dibangun dalam bahasa pemrograman Scala dan Java.

Apache Kafka bekerja dengan sistem *publish/subscribe messaging*. Dalam sistem ini dikenal istilah *Producer, Broker* dan *Consumer. Producer* merupakan suatu sistem yang mengambil data untuk topok tertentu secara *real-time*. Kemudian, topik tersebut didistribusikan kepada *Consumer* melalui *Broker. Consumer* merupakan sebuah sistem yang membutuhkan data untuk dikelola. Sebuah sistem pengolahan data dapat memiliki lebih dari satu *Consumer* dan *Producer.*

Untuk pengelolaan data dalam skala besar dengan menggunakan algoritma *machine learning,* kita dapat menggunakan Apache Spark. Apache Spark merupakan framework yang dapat melakukan analisa pada data dalam skala besar dan bersifat *opensource.* Untuk mendukung analisa data, Apache Spark dapat menampilkan banyak pilihan model visualisasi data.

Untuk mendukung kinerja Apache Kafka dan Apache Spark, kita dapat menggunakan Apache MesOS. Apache MesOS merupakan *cluster manager* yang memungkinkan kita untuk mengelola data secara terdistribusi.

2.4 Algoritma Rule Based *Classifier.*

Rule based classifier adalah algoritma klasifikasi yang menggunakan sekumpulan aturan IF-THEN untuk melakukan prediksi kelas pada suatu data. Secara sederhana, kita dapat mengekspresikan *Rule based classifier* dengan pseudocode di bawah ini:

IF kondisi THEN kesimpulan

Contoh penggunaan Rule Based Classifier dapat dilihat pada rule R1 di bawah ini:

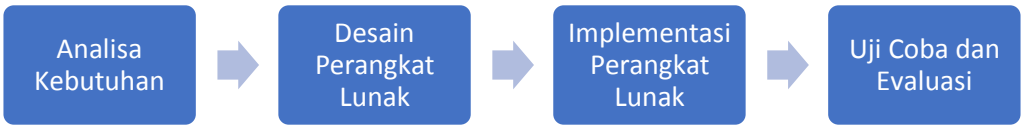
Rule 1

IF usia < 20 AND mahasiswa = yes THEN return yes;

Pada Rule based classifier terdapat prekondisi yang diwakilkan dengan pernyataan setelah IF, dan konsekuensi yang diwakilkan oleh pernyataan setelah THEN. Prekondisi dapat memiliki lebih dari satu syarat. Sementara konsekuensi berisi prediksi kelas dari data yang sedang diuji. Pada penelitian ini, Rule Based Classifier digunakan untuk menyeleksi data mana yang dibutuhkan oleh desainer kemasan dan data mana yang tidak dibutuhkan oleh desainer kemasan

3. METODOLOGI PENELITIAN

Metodologi yang digunakan dalam penelitian ini dapat dilihat pada Gambar 2. Metodologi tersebut terdiri dari empat tahap yaitu analisa kebutuhan, desain perangkat lunak, implementasi perangkat lunak serta uji coba dan evaluasi. Adapun penjelasan yang lebih rinci untuk setiap langkah dapat dilihat pada Subbab 3.1, dan



Subbab 3.2.

Gambar 2 Metodologi Pembuatan Perangkat Lunak

3.1 Analisa Kebutuhan

Mesin *Crawler* pada penelitian ini digunakan untuk mengambil data di media sosial sesuai dengan kriteria yang ditentukan oleh pengguna (desainer kemasan). Untuk

menjalankan tugas tersebut, mesin *crawler* setidaknya harus mampu melakukan dua hal yaitu mengunduh data serta menyeleksi data yang akan diunduh.

Data yang dikumpulkan oleh mesin *crawler* pada penelitian ini berasal dari dua media sosial yaitu Instagram dan Pinterest. Dari Pinterest, kita membutuhkan informasi terkait tren elemen grafis (tipografi, warna, bentuk dan ukuran). Untuk mendapatkan informasi tentang tipografi, ukuran dan bentuk, kita menggunakan data Note, Attribution dan Image. Sedangkan untuk mendapatkan tren warna, kita menggunakan data Color. Tabel 1 menunjukkan penjelasan lebih rinci tentang data yang akan diambil dari Pinterest.

Selain Pinterest, sumber data yang digunakan dalam penelitian ini adalah Instagram. Data yang dibutuhkan di Instagram berupa username konsumen pada segmen menengah ke atas beserta data turunannya yang berupa id, caption, media_count dan followers_count. Kriteria pengguna Instagram yang termasuk segmen konsumen menengah atas telah ditentukan sebelumnya oleh desainer kemasan**.** Tabel 2 menunjukkan keterangan lengkap tentang data yang diambil dari Instagram

Tabel 1 Data yang Diambil dari Pinterest

Data	**Keterangan**
Pin	Pin adalah istilah yang digunakan oleh Pinterest untuk merujuk item yang diunggah oleh pengguna
Note	Note memuat informasi tentang deskripsi dari Pin yang diunggah oleh pengguna.
Attribution	Attribution memuat informasi tentang metadata dari Pin yang meliputi informasi tentang pengunggah, judul serta tautan yang menjadi sumber gambar dari Pin tersebut.
Image	Image memuat data gambar yang ada dalam Pin
Color	Color memuat warna dominan dari gambar yang ada dalam Pin

Tabel 2 Data yang dibutuhkan dari Instagram

Data	**Keterangan**
Username	Username adalah identitas unik bagi setiap pengguna Instagram.
Id	Id yang digunakan dalam penelitian ini merujuk pada id media. Id media memuat identitas unik bagi setiap foto atau video yang diunggah oleh pengguna Instagram
Caption	Comments_count merujuk pada jumlah komentar bagi foto yang diunggah oleh pengguna Instagram
Media_count	Media_count merujuk pada jumlah foto atau video yang diunggah oleh pengguna Instagram
followers_count	Followers_count merujuk pada jumlah *follower* dari sebuah akun Instagram.

Untuk mendapatkan data yang telah disebutkan pada Tabel 1 dan Tabel 2, desainer kemasan harus mengisi input yang dibutuhkan oleh mesin *crawler.* Tabel 3 dan Tabel 4 menunjukkan input dan output dari mesin *crawler* yang akan dibuat.

Tabel 3 Skenario Input dan Output pada Pinterest

Input	Output
Username	Pin, Note , Atribution, Image, Color

Tabel 4 Skenario Input dan Output pada Instagram

Input	Output
Username	Daftar username follower dari username pengguna yang diinputkan dimana follower tersebut termasuk segmen konsumen kelas menengah atas.

3.2 Desain Arsitektur dan Implementasi Perangkat Lunak

Berdasarkan analisa kebutuhan pada Subab 3.1, kami mengembangkan algoritma dan arsitektur perangkat lunak yang bekerja di mesin *crawler.* Untuk membedakan segmen konsumen kelas menengah atas dengan segmen konsumen lain, desainer kemasan memberikan informasi berupa daftar produk yang biasa dikonsumsi oleh segmen menengah ke atas. Informasi ini nantinya digunakan sebagai bahan pertimbangan dalam melakukan seleksi data di Instagram. Proses seleksi data yang dikembangkan dalam penelitian ini dapat dibaca pada Subbab 3.2.1. dan Subbab 3.2.2. Adapun arsitektur mesin yang digunakan untuk mengambil data dan melakukan proses seleksi pada data tersebut dapat dilihat pada Subbab 3.2.3

3.2.1 Algoritma *Rule Based Classifier pada Instagram*

Untuk mendeteksi pengguna Instagram yang termasuk segmen kelas menengah ke atas, mesin *crawler* menggunakan algoritma *Rule Based Classifier.* Algoritma ini dilakukan dengan menyeleksi pengguna Instagram berdasarkan foto yang mereka unggah. Dalam foto yang ada di Instagram terdapat atribut caption. Desainer kemasan akan memberikan daftar nama produk yang biasa dikonsumsi oleh konsumen menengah ke atas. Sementara mesin *crawler* akan mengecek apakah caption dari foto yang diunggah pengguna Instagram memuat merk produk yang ada dalam daftar. Jika iya, maka mesin *crawler* akan mencatat username dari pengguna tersebut.

Skenario deteksi dimulai dengan memasukkan sebuah username dari pengguna Instagram. Selanjutnya, mesin *crawler* tersebut akan memeriksa satu persatu follower dari akun tersebut. Jika follower dari akun tersebut termasuk konsumen menengah atas, maka mesin *crawler* akan menyimpan username dari follower tersebut. Berikut ini *pseudocode* dari algoritma *Rule Based Classifier* untuk Instagram:

```
function: rule_based_classifier_Instagram (string username)
        segmen_menengah_atas[follower_count];
        for i=1 to i= follower_count
```

```
                if (follower[i]  == private account) then
                        go to next follower;
                        segmen_menengah_atas [follower] = 0;
                else
                        for j=1 to j= follower[i].media_count
                                caption = get follower[i].media[j].caption;
                                if (captionCheck(caption)==1) then
                                        count++;
                                        Go to next media;
                                end if
                        end for
                        if (count/follower[i].media_count >= 0.2) then
                                segmen_menengah_atas [follower] = 1;
                                Save username follower;
                        else
                                segmen_menengah_atas [follower] = 0;
                        end if
                end if
        end for
end function
```

Pada *pseudocode* di atas, terdapat fungsi CaptionCheck(). CaptionCheck() adalah sebuah fungsi boolean yang digunakan untuk mengecek apakah di dalam caption terdapat merk produk yang biasa dikonsumsi oleh konsumen menengah ke atas. Nilai 1 mewakili kondisi dimana caption yang diperiksa mengandung merk produk untuk konsumen menengah ke atas sementara nilai 0 mewakili kondisi yang sebaliknya.

3.2.2 Algoritma *Rule Based Classifier pada Pinterest*

Untuk mendeteksi Pin yang memuat informasi tentang kemasan, desainer kemasan akan menyiapkan kata kunci yang umumnya ada pada Notes dari Pin yang membahas tentang desain kemasan. Selanjutnya, mesin *crawler* akan memeriksa apakah Notes dari Pin yang sedang dituju mengandung kata kunci atau tidak. Jika iya, mesin *crawler* akan menyimpan gambar dan attribution dari Pin yang sedang diperiksa.

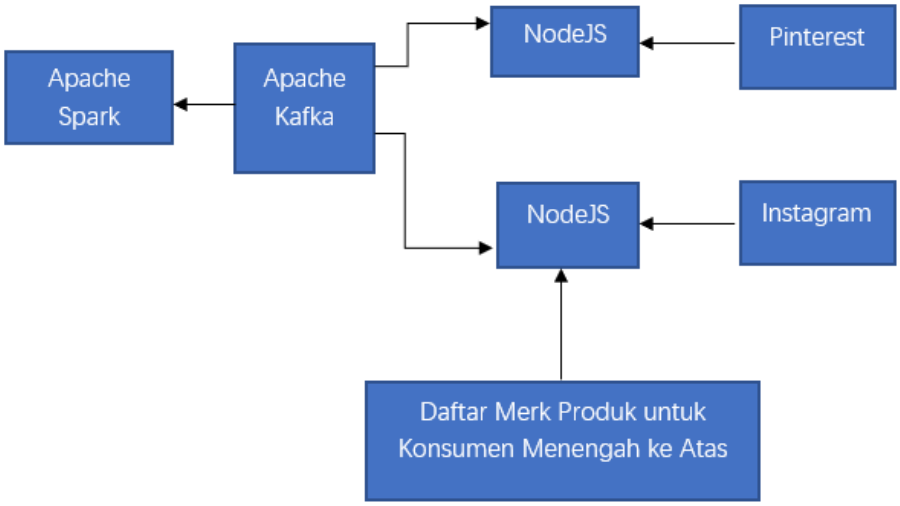
Adapun skenario untuk mengumpulkan data di Pinterest dimulai dengan desainer kemasan memberikan data input berupa username dari pengguna Pinterest. Kemudian, mesin *crawler* akan mengunjungi Board dan Pin dari pengguna tersebut untuk diperiksa satu per satu. Di bawah ini terdapat *pseudocode* dari algoritma yang digunakan oleh mesin *crawler* di Pinterest:

```
function: rule_based_classifier_Pinterest (string username)
                for i=1 to i=board_count
                        for j=1 to j=board[i].pin_count
                                string caps = get board[i].pin[j].notes;
                                if (caps contains keywords) then
                                        save board[i].pin[j].image;
                                        save board[i].pin[j].attribution;
                                end if
                        end for
```

end for
end function

3.2.3 Desain Arsitektur Mesin *Crawler*

Pada penelitian ini, mesin *crawler* dibangun dengan memanfaatkan Apache Spark, Apache Kafka dan NodeJS. Data yang dibutuhkan diambil melalui API pada Pinterest dan Instagram. Pengelolaan *streaming* ditangani oleh Apache Kafka. Sedangkan proses seleksi data dan visualisasi dikelola oleh Apache Spark. Pada arsitektur tersebut, terdapat *Rule Based Classifier* yang tersambung dengan API dari Instagram serta tersambung dengan basis data yang berisi merk produk untuk konsumen menengah ke atas. Gambar 3 menunjukkan desain arsitektur untuk membangun mesin *crawler.*



Gambar 3 Desain Arsitektur Mesin *Crawler*

4. UJI COBA DAN EVALUASI

Terdapat dua indikator untuk mengukur keberhasilan penelitian ini. Yang pertama adalah mesin *crawler* dapat mengumpulkan data dengan cepat. Yang kedua, data yang dikumpulkan oleh mesin *crawler* sesuai dengan kebutuhan pengguna. Kesesuian ini diukur dengan akurasi, presisi dan *recall*.

Untuk mengetahui apakah mesin *crawler* dalam penelitian ini memenuhi indikator keberhasilan atau tidak, kami menjalankan pengujian dalam dua skenario. Pada skenario pertama, kami menggunakan komputer *Single Node* yang ada pada Table 5. Adapun pada skenario kedua, kami menggunakan komputer *Multi Node* yang terdapat pada Tabel 6.

Tabel 5 Spesifikasi Komputer untuk Uji Coba dengan Skenario Single Node

Perangkat Keras	Perangkat Lunak
CPU: 8 Core, Memory : 8GB	CentOS 7 Server, Apache Kafka, Apache Spark, NodeJS, Scala, JDK 1.8

Tabel 6 Spesifikasi Komputer untuk Uji Coba dengan Skenario Multi Node

Perangkat Keras	Perangkat Lunak
MesOS Master (1 komputer) CPU : 8 Core Memory : 8 GB MesOS Agent (5 komputer) CPU : 5 x 8 Core Memory : 18 GB x 5	CentOS 7 Server, Apache Kafka, Apache Spark, Apache MesOS, HDFS, Scala, NodeJS, JDK 1.8

Skenario pengujian ini diterapkan pada 5 akun Instagram dan 5 akun Pinterest. Akun Instagram pertama hingga kelima masing-masing memiliki jumlah follower sebanyak 73.600, 366.900, 49.500, 13.000 dan 1.032.049. Sedangkan akun Pinterest pertama hingga kelima memiliki Pin masing-masing sejumlah 936, 6.129, 10.040, 75.171, dan 125.000. Tabel 7 menunjukkan akurasi, presisi dan recall yang didapatkan selama menjalani uji coba. Sedangkan Tabel 8 menunjukkan waktu eksekusi untuk masing-masing skenario.

Tabel 7 Presisi, *Recall* dan Akurasi pada Mesin *Crawler*

	Presisi (%)	Recall (%)	Akurasi (%)
Akun Instagram 1	71.4	45.5	60
Akun Instagram 2	88.9	47	50
Akun Instagram 3	71.6	52	50
Akun Instagram 4	34.8	48	56
Akun Instagram 5	90.9	64	62
Akun Pinterest 1	100	100	100
Akun Pinterest 2	100	100	100
Akun Pinterest 3	100	100	100
Akun Pinterest 4	100	100	100
Akun Pinterest 5	100	100	100

Tabel 8 Waktu Eksekusi Mesin *Crawler* pada Saat Uji Coba

	Waktu Eksekusi (detik)	
	Single Node	Multinode
Akun Instagram 1	6300	500
Akun Instagram 2	34300	3670
Akun Instagram 3	4100	1800
Akun Instagram 4	1400	480
Akun Instagram 5	110540	36500

	Waktu Eksekusi (detik)	
	Single Node	Multinode
Akun Pinterest 1	80	30
Akun Pinterest 2	580	110
Akun Pinterest 3	1100	230
Akun Pinterest 4	7400	2050
Akun Pinterest 5	13400	2600

Pada Tabel 7, kita dapat melihat presisi mesin *crawler* pada Instagram tidak stabil. Nilai tertingginya menyentuh angka 90% sedangkan nilai terendahnya hanya 34.8 %. Di samping itu, nilai akurasi dan *recall* mesin *crawler* masih di bawah 70%

Pada kasus ini, tidak stabilnya nilai presisi dipengaruhi oleh nilai *False Positive.* Pengguna Instagram yang termasuk pada daftar *False Positive* adalah toko online yang semua produknya biasa dikonsumsi oleh konsumen menengah ke atas. Kriteria seleksi yang digunakan pada algoritma *Rule Based* dalam penelitian ini mengenali konsumen menengah ke atas dengan menetapkan batas minimal jumlah produk yang diunggah di Instagram oleh seorang pengguna. Jika jumlah gambar produk untuk segmen menengah atas yang diunggah melebihi batas minimal, maka pengguna tersebut termasuk dalam segmen konsumen menengah atas. Pada kasus toko online, semua gambar yang diunggah adalah gambar produk untuk konsumen menengah atas. Jadi dapat dipastikan bahwa semua toko online lolos dari seleksi.

Rendahnya nilai *recall* menunjukkan kurangnya kemampuan mesin dalam mengenali segmen konsumen menengah atas dengan tepat. Mesin ini mengenali produk untuk segmen konsumen menengah atas dengan memeriksa *caption.* Sementara itu, tidak semua pengguna Instagram menuliskan merek produk di *caption.*

Berdasarkan hasil uji coba ini, kita dapat menentukan langkah perbaikan selanjutnya. Jumlah *False Positive* dapat diturunkan dengan merumuskan kriteria yang layak tentang bagaimana sebuah akun pengguna Instagram dapat dikenali sebagai toko online. Sedangkan jumlah *False Negative* dapat diturunkan dengan mengembangkan algoritma pengenalan gambar untuk mendeteksi merk pada gambar yang diunggah oleh pengguna Instagram

Hal lain yang dapat kita lihat pada Tabel 7 adalah nilai akurasi mesin *crawler* pada Pinterest yang sampai 100%. Tingginya nilai akurasi ini disebabkan oleh perilaku para pengguna Pinterest yang menuliskan keterangan lengkap pada gambar yang mereka unggah. Hal ini membuat mesin *crawler* dapat dengan mudah menemukan data yang dituju.

Pada Tabel 8, kita dapat melihat bahwa penggunaan komputer multinode berpengaruh pada waktu eksekusi mesin *crawler.* Akun Instagram dengan 1.032.049 *follower* dapat diproses dalam waktu 36500 detik atau satu jam. Dimana sebelum menggunakan komputer *multi node,* pengolahan untuk akun Instagram ini membutuhkan waktu lebih dari 30 jam.

5. KESIMPULAN

Berdasarkan hasil uji coba yang telah dilakukan, mesin *crawler* yang dibangun pada penelitian ini memiliki akurasi, recall dan presisi yang masih perlu ditingkatkan lagi. Sedangkan dari sisi arsitektur perangkat lunak, kombinasi antara *Apache Kafka, Apache Spark* dan *Apache MesOS* dapat membuat proses *streaming* dan seleksi data berjalan dengan cepat. Adanya *Apache MesOS* memungkinkan pemrosesan data berjalan secara terdistribusi di banyak node. Hal ini dapat mempercepat pengolahan data secara signifikan.

DAFTAR PUSTAKA

Clement, J., Kristensen, T., & Gronhaug, K. (2013). Understanding Consumer's In-Store Visual Perception: The Influence of Package Design Features on Visual Attention. *Journal of Retailing and Consumer Services*, 234-239.

Graves, P. (2010). *Consumerology : The Market Research Myth, The Truth about Consumers and The Psychology of Shopping.* London: Nicholas Braley Publishing.

Kominfo, K. (2017, Mei 31). *Kominfo : Pengguna Internet di Indonesia 63 Juta Orang.* Retrieved from Kementrian Komunikasi dan Informasi: https://kominfo.go.id/index.php/content/detail/3415/Kominfo %3A Pengguna Internet di Indonesia 63 Orang/0/berita_satker

Kotler, P., & Keller, K. (2011). *Marketing Management .* New Jersey: Prentice Hall.

Kusrianto, A. (2007). *Packaging Design*. Majalah Concept.

Majalah Marketing. (2016, Desember). *12 Karakter Unik Konsumen Indonesia*. Indonesia: Majalah Marketing.

Pickton, D., & Broderick, A. (2005). *Integrated Marketing Communication .* Prentice Hall.

Pine, J., & Gilmore, J. (2011). *The Experience Economy, Update Edition.* Harvard Business Press.

Rundh, B. (2009). Packaging Design: Creating Competitive Advantage with Product Packaging. *British Food Journal 111 (9)*, 988-1002.

Sowardikoen, W. (2013). *Metodologi Penelitian Visual.* Bandung: Dinamika Komunika.

Vyas, H. (2015). Packaging Design Elements and Users Perception: A Context in Fashion Branding and Communication. *Journal of Applied Packaging Researches Vol 7 No 2*.

Zafarani, R., Abbasi, A., & Huan, L. (2014). *An Introduction To Social Media Mining.* Cambridgeshire: Cambridge University Press.