BAB II
TINJAUAN PUSTAKA

2.1 Scraper/Crawler

Web crawler atau yang dikenal juga dengan istilah *web spider* atau *web* robot adalah program yang bekerja dengan metode tertentu dan secara otomatis mengumpulkan semua informasi yang ada dalam suatu website. *Web crawler*, yang sering disebut *crowler* saja, akan mengunjungi setiap alamat *website* yang diberikan kepadanya, kemudian mengorek, mengambil, dan menyimpan semua informasi yang terdapat didalam *website* tersebut (Wijaya, 2019). Dengan memindahkan isi sebuah *website* ke dalam komputer lokal, kita bisa menelusuri isi website tersebut tanpa harus terhubung ke internet. Tujuannya adalah untuk mempermudah analisis srtuktur sebuah *website* secara *offline*. Namun, perlu Anda ketahui, bahwa tidak semua isi website bisa dipindahkan ke dalam komputer lokal. Salah satunya adalah *website* yang terbuat dari flash tidak bisa dipindahkan secara sempurna karena link yang terdapat di dalamnya tidak tersimpan dalam *file* HTML maupun script PHP (Zam, 2011).

Proses *web crawler* dalam mengunjungi setiap dokumen *web* disebut dengan *web crawling* atau *spidering*. Proses *crawling* dalam suatu *website* dimulai dari mendata seluruh url dari *website*, menelusurinya satu-persatu, kemudian memasukkannya dalam daftar halaman pada indeks *search engine*, sehingga setiap kali ada perubahan pada website, akan terupdate secara otomatis. Web crawling adalah proses mengambil kumpulan halaman dari sebuah web untuk dilakukan pengindeksan (*indexing*) untuk mendukung kinerja mesin pencari.

Web crawler biasa digunakan untuk membuat salinan secara sebagian atau keseluruhan halaman *web* yang telah dikunjunginya agar dapat diproses lebih lanjut oleh sistem penyusun index. *Crawler* dapat juga digunakan untuk proses pemeliharaan sebuah *website*, seperti memvalidasi kode html sebuah *web*, dan *crawler* juga digunakan untuk memperoleh data yang khusus seperti mengumpulkan alamat *e-mail*.

Web crawler termasuk kedalam bagian *software agent* atau yang lebih dikenal dengan istilah program *bot*. Secara umum *crawler* memulai prosesnya dengan memberikan daftar sejumlah alamat website untuk dikunjungi, disebut sebagal seeds. Setiap kali sebuah halaman web dikunjungi, *crawler* akan mencari alamat yang lain yang terdapat didalamnya dan menambahkan kedalam daftar *seeds* sebelumnya.

2.2 Data Mining

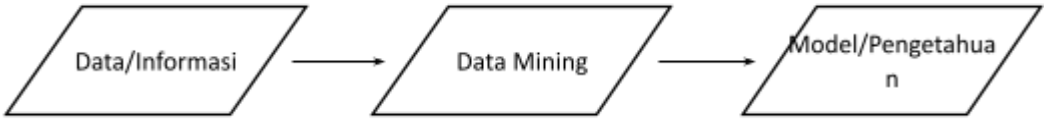
Data mining merupakan proses penggalian informasi dan pola yang bermanfaat dari data yang sangat besar. Data mining mencakup pengumpulan data, ekstraksi data, analisis data, dan statistik data. Data mining juga dikenal sebagai *knowledge discovery*, *knowledge extraction*, data/*pattern analysis*, *information harvesting*, dan lain-lain (Muflikhah and Ratnawati, 2018).

Data mining juga merupakan proses logis untuk menemukan informasi yang berguna. Setelah ditemukan informasi dan pola dapat digunakan untuk alat pendukung dalam pengambilan keputusan dalam mengembangkan bisnis. Alat data mining dapat memberikan jawaban untuk berbagai pertanyaan yang terkait dengan bisnis dan terlalu sulit untuk diselesaikan. *Data mining* juga dapat digunakan untuk meramalkan tren masa depan yang memungkinkan pebisnis membuat keputusan yang efektif, proaktif, dan dinamis.

Definisi lainnya untuk data mining adalah suatu proses menganalisis pola data yang tersembunyi menurut berbagai perspektif untuk kategorisasi menjadi informasi yang berguna, yang dikumpulkan di area umum, data warehouse untuk analisis yang efisien, algoritma data mining, memfasilitasi pengambilan keputusan bisnis, dan informasi lainnya. Data mining menggunakan analisis matematika dalam mendapat atau menemukan pola dan kecenderungan dari data. Pada umumnya, pola ini sukar ditemukan oleh eksplorasi data secara biasa/tradisional, hal ini disebabkan oleh terlalu rumitnya hubungan antardata atau juga dapat disebabkan oleh data yang begitu besar.

Data mining bertujuan untuk menemukan pola yang sebelumnya tidak diketahui. Jika pola-pola tersebut telah diperoleh maka dapat digunakan untuk menyelesaikan berbagai macam permasalahan. Data mining saat ini juga telah

menjadi suatu teknologi baru yang kuat dengan potensi besar untuk membantu perusahaan fokus pada informasi paling penting dalam data yang telah mereka kumpulkan tentang perilaku pelanggan dan pelanggan potensial mereka. Melalui *data mining*, perusahaan dapat menemukan informasi dalam data yang begitu besar melalui pengolahan yang tepat dan efektif dengan berbagai metode yang ada dalam *data mining* sehingga secara sederhana data mining dapat digambarkan sebagai suatu pola atau model atau kaidah atau pengetahuan yang dihasilkan dari data mining, seperti Gambar 2.1.



Gambar 2.1 Model atau pengetahuan Merupakan Output Data Mining

Untuk menghasilkan model ataupun pengetahuan maka sangat berpengaruh dari bagaimana data atau informasi sebelumnya atau bagaimana sejarah data atau informasi terdahulu sehingga dapat diprediksi atau diestimasi informasi di masa yang akan datang. Makin bagus informasi masa lampau atau data-data sebelumnya maka biasanya hasil yang diperoleh memiliki keakuratan yang sangat baik.

Sebagai contoh, jika seorang manager supermarket ingin mengetahui apakah seseorang pelanggan ketika membeli telur apakah dia juga akan membeli minyak? Prosesnya berdasarkan histori datanya dapat dilihat berikut ini (Arhami and Nasir, 2020).

Historical Data

BaskedId	***Eggs***	***Oil***	***Milk***	...
1	*yes*	*yes*	*no*	...
2	*no*	*yes*	*yes*	...
3	*no*	*no*	*yes*	...
4	*no*	*yes*	*yes*	...
5	*yes*	*yes*	*no*	...
6	*yes*	*no*	*yes*	...
7	*no*	*no*	*no*	...
8	*yes*	*yes*	*yes*	...
...	...	...	...	...

Pola/Model: Eggs ---> Oil: Confodence = 75%, Support 37%

2.3 Machine Learning

Machine Learning adalah ilmu yang mempelajari tentang algoritma komputer yang bisa mengenali pola-pola di dalam data, dengan tujuan untuk mengubah beragam macam data menjadi suatu tindakan yang nyata dengan sesedikit mungkin campur tangan manusia. Dengan *Machine Learning*, kita dapat menciptakan mesin (komputer) yang "belajar" dari data yang ada, selanjutnya dia bisa membuat keputusan secara mandiri tanpa perlu diprogram lagi. Secara umum, *Machine Learning* berada di bawah payung *Artificial Intelligence*/AI, (kecerdasan buatan) (Kurniawan, 2020).

2.4 **Pentaho**

Pentaho Data Integration atau Kettle adalah *tools* yang memiliki kemampuan *extract*, *transform*, dan *load* (ETL) pada multi *platform database*. *Script* dari disain dapat disimpan dalam bentuk *file* ataupun *repository*. Selain itu, pada *tools* ini terdapat cukup banyak *'steps'* untuk mengatur *workflow control* (JOB), dan data *workflow* (*Transformation*) (Subarkah *et al.,* 2022).

Di dalam pembahasa PDI, akan muncul beberapa istilah, antara lain:

a. *Extract*: Proses pengambilan data dari *datasource*
b. *Transform*: Proses pengubahan data yang telah *diextract*
c. *Load*: Proses *store*/penyimpanan data yang telah *ditransform*
d. *Job*: *File* yang berekstensi .kjb yang berfungsi sebagai *workflow* control
e. *Transformation*: *File* yang berekstensi .ktr yang berfungsi sebagai data *workflow*. *Kettle*: Nama lain dari *Pentaho Data Integration*
f. *Spoon*: Aplikasi GUI untuk merancang atau menjalankan *job*/*transformation*
g. *Pan*: Utilitas untuk menjalanan *transformation* dalam tampilan *console*. Biasanya digunakan untuk otomasi terjadwal

Server BI yang berjalan sebagai *web application portal* yang terdiri dari layanan *web service, workflow* pada *space JVM (Jasa Virtual Machine)*, dan sebagai *user interface* untuk laporan operasional maupun analisis. *Workflow*

berupa integrasi produk Pentaho yang telah disebutkan sebelumnya (*Pentaho Data Integration, Pentaho Reporting,* dan *Pentaho Analysis*) dalam bentuk *solution*. *Script JVP (Java Server Pages)* dapat dengan mudah diintegrasikan ke dalam *platform*. User dapat memiliki *space* sendiri untuk menyimpan *report* dan tipe *solution* lainnya. *Ad hoc report* yang bisa digunakan untuk membuat rancangan *report on the fly* tanpa keterlibatan IT.

2.5 Python

Code Python adalah bahasa pemrograman interpretatif yang bisa dipasang pada berbagai *platform*, khususnya *platform* yang berfokus pada keterbacaan kode. Kode *Python* bisa di-*embed* ke bahasa lain seperti C dan Java, atau sebaliknya, dari bahasa C atau Java ke *Python*.

Pemrograman *Python* itu merupakan salah satu bahasa pemrograman yang dapat melakukan eksekusi sejumlah instruksi multu guna secara langsung dengan metode *Object Oriented Programming* dan menggunakan sergantik dinamis untuk memberikan tingkat keterbacaan sintak. *Pemrograman Python* memiliki bahsa yang kemampuan, menggabungkan kapabilitas dan sintaksis kode yang jelas dan dilengkapi dengan fungsionalitas pustaka standar yang besar serta komprehensif. Pemrograman bahasa *Python* difokuskan untuk digunakan dalam menganalisis data, visualisasi data, membuat dan mengembangkan Al. Pemrograman *Python* adalah pemrograman yang paling mudah di pelajari dengan *code* yang pendek dan tidak susah. *Python* memiliki pustaka (*library*) yang luas dan dapat dikembangkan ke bidang-bidang lainnya. Beberapa *library python* yang popular *dalam Data Science* dan Al adalah *Scikit- Learn,TensorFlow,PyTorch*.

Pemrograman bahasa *Python* merupakan bahasa pemrograman yang tidak menggunakan *compiler*. Dengan sifat *open-source* yang dimilikinya, pengguna dapat mempelajari *Python* dengan mudah karena bahasa ini dapat dugunakan dalam membuat situs.mengembangkan situs, mengembangkan *video game*, membangun *GUI Desktop*, dan mengembangkan perangkat lunak (Setiawan and Vania, 2022).

2.6 Penelitian Terkait

Penelitian yang digunakan terkait dengan penelitian saat ini adalah seperti tabel 2.1.

Tabel 2.1 Penelitian Terkait

No	Judul Penelitian	*Tools*	Kesimpulan
1	Implementasi Data Warehouse Dan Penerapannya Pada Toko Magnifique Clothes Dengan Menggunakan Tools Pentaho (Subuh and Yasman, 2019)	Pentaho	Pemilik toko dalam melihat perkembangan keuntungan dan penjualan yang terjadi setiap minggu, bulan dan tahun, sehingga pemilik dapat melakukan analisis terhadap penyampaian informasi yang sudah disajikan dalam bentuk grafik atau dashboard
2	Analisis Sentimen Calon Presiden Indonesia 2019 Berdasarkan Komentar Publik di Facebook (Santoso and Nugroho, 2019)	Web crawling	Hasil dari penelitian berdasarkan data yang telah dikumpulkan sejak tanggal 17 April 2019 sampai 22 Mei 2019, Joko Widodo lebih unggul polaritas sentimen positif dari data sebanyak 5.000 komentar yang dipilih secara acak pada masing-masing calon presiden dan melalui tahap preprocessing yang menghasilkan polaritas sentimen. Joko Widodo diperoleh 85% sentimen positif, dan 15% untuk sentimen negatif. Sedangkan Prabowo Subianto diperoleh

No	Judul Penelitian	*Tools*	Kesimpulan
			76% sentimen positif, dan 24% untuk sentimen negatif
3	Seleksi Fitur Support Vector Machine pada Analisis Sentimen Keberlanjutan Pembelajaran Daring (Natasuwarna, 2020)	Web crawling	Data yang digunakan berjumlah 200 data tweet terdiri dari 100 komentar positif dan 100 komentar negatif menggunakan lima rasio perbandingan data latih dan data uji. Penelitian menghasilkan evaluasi tertinggi pada pada 8-Fold Cross Validation dengan accuracy sebesar 86,00%, precision sebesar 87,38%, dan recall sebesar 85,02%
4	Membangun Web Crawler Berbasis Web Service untuk Data Crawling Pada Website Google Play Store (Ilmawan, 2018)	Web Crawler	Sistem yang dibangun berhasil mengambil data pada website Google Play Store dengan baik dan benar sesuai dengan kebutuhan pada analisis sistem dan dapat diintegrasikan dengan web service berbasis REST untuk mendukung penggunaan sistem secara cross platform
5	Analisis Sentimen Masyarakat Indonesia Terhadap Pemindahan Ibu Kota Negara Indonesia pada Twitter (Lestari, Mupaat and Erfina, 2022)	Support Vector Machine (SVM)	Dari hasil pemprosesan data terhadap komentar yang terdapat di twitter terkait dengan pemindahan Ibu Kota Negara Indonesia dan penggunaan nama Nusantara, maka didapatkan 1.141 komentar positif dan 591 komentar negatif. Hal ini menunjukkan bahwa masyarakat Indonesia yang beranggapan

No	Judul Penelitian	*Tools*	Kesimpulan
			positif terhadap Ibu Kota Negara baru Indonesia
6	Analisis Sosial Media Pemerintah Daerah di Indonesia Berdasarkan Respons Warganet (Furqon *et al.*, 2018)	Web Crawler	Tingkat sentimen pada halaman Facebook pemerintah daerah di Indonesia masih tergolong positif atau dengan kata lain terdapat respons positif dari masyarakat terhadap isi posting tersebut
7	Ekstraksi Data Produk E-Marketplace Sebagai Strategi Pengolahan Segmentasi Pasar Menggunakan Web Crawler (Surahman, Octaviansyah and Darwis, 2020)	Web Crawler	Kesuksesan informasi yang disajikan dalam bentuk segmentasi pasar sebesar 79 % dan nilai tersebut memiliki presentase tanggapan responden dalam kriteria Baik