BAB 2
LANDASAN TEORI

2.1 Content Based Filtering

Content-based filtering merupakan salah satu teknik *Machine Learning* yang memakai *attribute similarity* untuk membuat keputusan. Teknik ini biasa digunakan dalam pembangunan sistem yang memberikan sebuah rekomendasi, yaitu rancangan algoritma untuk mengiklankan/merekomendasikan sesuatu kepada *user* berdasarkan data yang terkumpul tentang *user* [13]. Metode ini menghasilkan sebuah rekomendasi dengan menggunakan *keyword* dan atribut yang ditetapkan ke objek di dalam database dan mencocokkannya dengan profil *user*. Profil *user* dibuat berdasarkan data yang diperoleh dari aktivitas user, seperti penilaian (suka dan tidak suka) atau item yang dicari di situs website [15]. Metode ini digunakan untuk merekomendasikan item kepada pengguna berdasarkan preferensi dan minat mereka sebelumnya. Ini menggunakan konten item (seperti deskripsi film, ringkasan buku, dll.) untuk merekomendasikan item serupa. Cara ini sering digunakan di situs belanja online dan layanan streaming seperti Netflix dan Amazon. Sistem rekomendasi yang menggunakan metode *content-based filtering* akan memberikan hasil rekomendasi item yang memiliki *similarity* dengan item yang dipilih atau disukai *user* [16]. Keuntungan model ini tidak memerlukan data apa pun tentang *user* lain, karena rekomendasinya khusus untuk *user* ini. Model ini dapat meningkatkan akurasi hasil rekomendasi, model ini juga memiliki kemampuan untuk membuat rekomendasi yang lebih spesifik, dan juga memiliki kemampuan untuk membuat rekomendasi berdasarkan preferensi *user* [14].

2.2 Anime

Anime adalah sebuah kata serapan bahasa Inggris di dalam bahasa Jepang dari kata *"Animation"*. Seiring dengan berkembangnya zaman, anime telah masuk menjadi sebuah kategori di dalam daftar film [6]. Dengan grafik warna-warni, karakter yang hidup, dan tema yang fantastis menjadi sebuah tanda bahwa anime adalah gaya visual untuk animasi yang berasal dari Jepang. Anime sering didasarkan pada sebuah manga, atau buku komik Jepang, dan telah menjadi populer di seluruh dunia. Kata anime berasal dari kata *animation* dalam bahasa Inggris yang disingkatkan dan merujuk kepada semua tipe anime [17]. Pada awalnya anime

hanya ada di televisi tetapi seiring dengan berkembangnya teknologi, anime bisa diakses dengan mengakses aplikasi atau *website* jasa *streaming* anime.

2.3 Web Scraping

Web Scraping merupakan teknik mengambil informasi atau data dari suatu *website* dengan cara memanfaatkan struktur *HTML* atau*XML* dari website tersebut. Proses dari teknik ini biasanya dilakukan dengaan menggunakan sebuah program atau sebuah *code* yang bisa mengambil data dari *website* secara otomatis. *Web Scraping* adalah salah satu macam dari *data mining*. Yang menjadi tujuan penting dari langkah *Web Scraping* adalah untuk mendapatkan informasi yang masih tidak terstruktur datanya dari *website* dan mengubahnya menjadi struktur agar nantinya bisa dipahami dengan lebih mudah seperti *spreadsheet, database* atau *file comma-separated values (CSV). Web scrapping* sering digunakan untuk mengumpulkan data yang diperlukan untuk analisis, penelitian, atau keperluan lainnya dari sebuah atau sejumlah *website*. Meskipun *web scrapping* dapat memberikan banyak manfaat, terutama dalam mengumpulkan data yang diperlukan, ada juga beberapa pertimbangan etis yang perlu diperhatikan. Sebagian *website* mungkin tidak mengizinkan proses *web scrapping*, sehingga perlu diperhatikan untuk tidak melakukan *web scrapping* secara tidak sah atau melanggar hak cipta dari *website* tersebut [18].

2.4 Preprocessing

Preprocessing adalah proses menyiapkan data untuk dianalisis dengan membersihkan, mengubah, dan mengaturnya. Ini termasuk tugas-tugas seperti menghapus outlier, menormalkan data, dan menyandikan variabel kategori. Preprocessing adalah langkah penting dalam alur kerja ilmu data karena membantu memastikan bahwa data siap untuk dianalisis. Data yang sudah melalui tahap-tahap *preprocessing* akan menjadi data yang lebih tersktruktur [19]. Tahap-tahap *preprocessing* sebagai berikut:

1. *Case Folding* merupakan proses mentransformasi semua huruf dan kata di dalam data anime baik itu judul, genre ataupun studio anime di dalam dokumen menjadi huruf kecil. Ini membantu mengurangi *size* kosakata dan meningkatkan akurasi algoritma klasifikasi teks [20].

2. Tokenisasi merupakan proses memecah text menjadi *words, phrases, symbols* atau elemen lainnya yag disebut token [20].

3. Eliminasi merupakan teknik dari *preprocessing* yang digunakan untuk mengurangi jumlah fitur dalam sebuah *dataset* dan menghapus *duplicated words*. Kata yang *duplicated* diasumsikan memiliki fitur yang sama, hanya akan disimpan 1 kata jika ada kata yang sama [21].

4. *Filtering* merupakan Teknik *preprocessing* yang melibatkan subset data dari dataset asli berdasarkan kriteria tertentu. *Filtering* ini dapat digunakan untuk mengurangi *size dataset*, menghapus data yang tidak relevan atau fokus pada fitur tertentu [19].

5. *Stemming* merupakan teknik *preprocessing* yang digunakan untuk mengurangi jumlah kata di dalam dokumen dengan menghilangkan *prefix & Suffix* dengan kata lain menstransformasikan suatu kata yang memiliki awalan ataupun akhiran menjadi hanya kata dasar [20].

2.5 Cosine Similarity

Di dalam *data mining*, ukuran kemiripan mengacu pada jarak dengan dimensi yang mewakili fitur objek data di dalam kumpulan data. Jika jarak ini lebih kecil maka tingkat kemiripannya akan tinggi, tetapi jika jaraknya besar maka tingkat kemiripannya akan rendah. *Cosine Similarity* adalah kosinus sudut antar vektor. Vektor biasanya bukan nol dan berada dalam ruang hasil kali dalam [22]. *Cosine Similarity*

$$\text{sim(A, B)} = \frac{n(A \cap B)}{\sqrt{n(B)n(B)}} \tag{2.1}$$

sudut antara dua vektor dan biasanya digunakan untuk menghitung kemiripan antara dua objek [23]. Fungsi cosine similarity antara item A dan item B ditunjukkan sebagai berikut.

Keterangan:

$sim(A,B)$ = nilai kemiripan dari item A dan item B.
$n(A)$ = banyaknya fitur pada konten item A.
$n(B)$ = banyaknya fitur pada konten item B.
$n(A,B)$ = banyaknya fitur konten yang terdapat pada item A dan pada item B.

Kedua objek yang memiliki hasil nilai *similarity* sama dengan 1 atau semakin besar nilai fungsi *similaritynya, kedua objek yang dilakukan penghitungan dianggap serupa atau bisa identik dan sebaliknya*

2.6 Top-N Recommendation

Top-N recommendation adalah sebuah teknik yang digunakan dalam sistem rekomendasi untuk menyarankan sejumlah item terbaik kepada *user*. Nilai-nilai hasil perhitungan *cosine similarity* digunakan untuk memberikan *rank* rekomendasi kepada *user*. Nilai hasil perhitungan *cosine similarity* dengan nilai *similarity* lebih tinggi diprediksi akan menjadi pilihan *user*. Untuk menentukan *item-item* terbaik yang akan disarankan kepada pengguna digunakanlah metode *filtering* [24].

2.7 Confusion Matrix

Confusion Matrix adalah metode yang populer dipakai saat memecahkan masalah klasifikasi dan bisa dipakai untuk mengetahui kinerja suatu sistem dengan membandingkan hasil klasifikasi dari sistem dengan klasifikasi sebenarnya. Ini dapat diterapkan untuk klasifikasi biner serta untuk masalah klasifikasi multikelas. Contoh matriks konfusi untuk klasifikasi biner ditunjukkan pada Tabel 2.1 [25] [26].

Tabel 2.1. *Confusion Matrix*

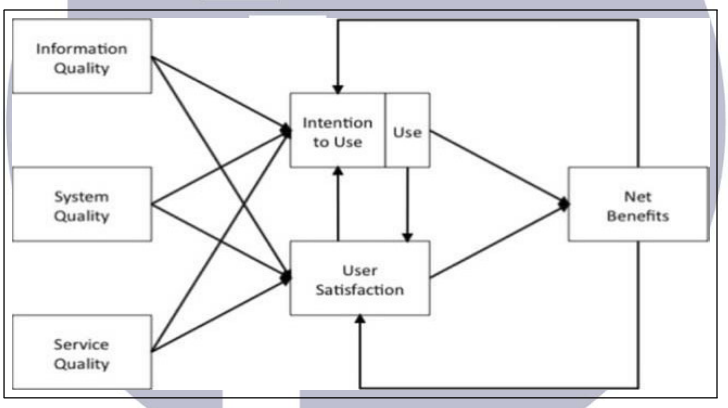
		Nilai Sebenarnya	
		TRUE	FALSE
Nilai Prediksi	TRUE	True Positive (TP)	False Positive (FP)
	FALSE	False Negative (FN)	True Negative (TN)

$$\text{presisi} = \frac{TP}{TP+FP} \tag{2.2}$$

$$\text{akurasi} = \frac{TP+TN)}{TP+TN+FP+FN} \tag{2.3}$$

2.8 Model Delone dan Mclean

Model Delone & Mclean adalah model untuk menentukan keberhasilan sistem informasi yang dikembangkan oleh DeLone dan McLean pada tahun 1992. Hal ini didasarkan pada premis bahwa keberhasilan sistem informasi merupakan fungsi dari lima dimensi utama: *System Quality, Information Quality, Service Quality, User Satisfaction and Net Benefit* [27].



Gambar 2.1. Model Delone dan Mclean