BAB 2 LANDASAN KEPUSTAKAAN

2.1 Kajian Pustaka

Terdapat suatu artikel ilmiah yang dibuat oleh Kim Schouten dan Flavius Frasincar (2015) tentang perkembangan analisis sentimen untuk tingkat aspek. Bidang analisis sentimen merupakan gabungan dari pemrosesan bahasa alami, sistem temu kembali informasi, dan kecerdasan buatan (Schouten dan Frasincar, 2015). Secara umum tahapan dalam analisis sentimen tingkat aspek terdiri dari tiga proses utama, yaitu identifikasi, klasifikasi, dan pengumpulan atau perangkuman. Namun tidak semua penelitian menggunakan semua proses tersebut atau dengan urutan yang sama. Untuk pendekatan yang biasa digunakan dalam analisis sentimen, secara garis besar bisa dibagi ke dalam tiga kategori, yaitu deteksi atau identifikasi aspek, analisis sentimen, penggabungan deteksi aspek dan analisis sentimen. Kemudian, dilihat dari karakteristik algoritme yang digunakan pada deteksi atau identifikasi aspek, diantaranya ada *Frequency-Based, Syntax-Based, Supervised Machine Learning, Unsupervised Machine Learning,* dan *Hybrid.* Untuk analisis sentimen, terdapat *Dictionary-Based, Supervised Machine Learning, Unsupervised Machine Learning.* Untuk penggabungan deteksi aspek dan analisis sentimen, ada *Syntax-Based, Supervised Machine Learning, Unsupervised Machine Learning, Hybrid Machine Learning.*

Untuk topik tentang sentimen analisis pada tingkat aspek ini, sudah ada beberapa penelitian sebelumnya dengan berbagai metodologi, salah satunya adalah dari Mohsen Farhadloo dan Erik Rolland (2013) dengan judul "*Multi-Class Sentiment Analysis with Clustering and Score Representationn".* Pada penelitian tersebut terdapat tiga hasil klasifikasi untuk sentimen, yaitu positif, netral, dan negatif. Untuk data yang digunakan diambil dari ulasan di TripAdvisor. Data tersebut terdiri dari 2.405 kalimat ulasan dengan rincian 992 ulasan positif, 992 ulasan netral, dan 421 ulasan negatif. Pada penelitian tersebut, peneliti mengusulkan penggunaan *Bag of Noun*s untuk proses klasterisasi. Pada tahapan klasterisasi tersebut ditujukan untuk mendapatkan aspek yang banyak diulas. Hasilnya mampu meningkatkan kinerja klasterisasi dan menjadikan algoritme tersebut efektif untuk identifikasi aspek (Farhadloo dan Rolland, 2013). Selain *Bag of Nouns,* peneliti juga mengusulkan ekstraksi fitur dengan *score representation* sebagai fitur yang akan digunakan dalam pengklasifikasian. Skor tersebut terdiri positif, netral, dan negatif. Hasilnya menunjukkan peningkatan kinerja untuk analisis sentimen dengan tigas kelas sebesar 20% menjadi 69% berdasarkan *average f1-score* dibandingkan dengan penelitian sebelumnya.

Selanjutnya penelitian dari Anuj Sharma dan Shubhamoy Dey (2013) tentang penggunaan jaringan saraf tiruan untuk analisis sentimen. Algoritme jaringan saraf tiruan yang digunakan adalah SOM dan LVQ dengan optimasi. SOM termasuk dalam algoritme *unseprvised,* sedangkan LVQ termasuk dalam algoritme *supervised*. Data yang digunakan adalah ulasan film yang terdiri dari 1.000 ulasan positif dan 1.000 ulasan negatif. Hasil percobaan menunjukkan algoritme

supervised memiliki kinerja yang lebih baik untuk analisis sentimen dibanding algoritme *unsupervised*. Hasil percobaan juga menunjukkan bahwa dengan *multi-pass* memberikan hasil yang lebih baik dibanding *single-pass.*

Dalam penelitian yang dilakukan oleh Zhexue Huang (1998), K-Means akan bekerja untuk nilai numerik namun untuk nilai kategorikal diperlukan model perhitungan yang berbeda sehingga diajukanlah K-Modes sebagai K-Means yang telah dimodifikasi.

Dalam penelitian yang dilakukan oleh Jamadar dan Kakade (2015), LVQ dapat menyaingi algoritme SVM pada studi kasus deteksi tumor otak dengan memberikan akurasi yang lebih tinggi. Hasil dari SVM diperoleh dari klasifikasi yang tidak intuitif (Hammer, Strickert, dan Villmann, 2004) yang mana memerlukan data latih yang besar untuk hasil yang optimal. Sebagai alternatif dari SVM, terdapat algoritme LVQ yang merupakan klasifikasi yang intuitif karena LVQ merupakan salah satu algoritme jaringan saraf tiruan yang meniru cara kerja sistem saraf yang ada di otak kita sehingga dia dapat belajar. LVQ sendiri terdapat beberapa versi salah satunya yang merupakan modifikasi pertama adalah LVQ2. Berdasarkan penelitian dari Budianita dan Firdaus (2016) tentang penggunaan LVQ2 pada diagnosis penyakit jiwa menunjukkan hasil dari LVQ2 yang memiliki *window* dapat meningkatkan nilai akurasi dari LVQ dasar yang tanpa *window.*

Dalam penelitian yang dilakukan oleh Thomas Villmann, Andrea Bohnsack, dan Marika Kaden (2017) tentang penggunaan LVQ sebagai alternatif dari SVM. Penggunaan LVQ sebagai alternatif dari SVM disini adalah dengan pertimbangan bahwa LVQ lebih mudah untuk dipahami disaat SVM yang sering susah untuk dijelaskan dan memerlukan pengethuan teoritis yang tinggi untuk penggunaannya (Villmann, T., Bohnsack, A., dan Kaden, M., 2017). Seperti hasil penelitian yang dilakukan oleh Jamadar dan Kakade (2015) pada deteksi dan analisis tumor otak dengan membandingkan penggunaan SVM dan LVQ, hasilnya dengan LVQ memberikan akurasi 90% lebih tinggi daripada SVM yang memiliki akurasi 86,67%.

2.2 Profil HARRIS Hotel & Conventions Malang

HARRIS Hotel merupakan salah satu hotel terkemuka di Indonesia. Dilansir dari situs TripAdvisor, HARRIS Hotel & Conventions Malang yang merupakan hotel bintang 4 ini terpilih sebagai “Hotel Terbaik untuk Keluarga - Indonesia” dengan peringkat 19 dari 25. Hotel ini juga mendapatkan peringkat 4 dari 56 Hotel di Malang dengan penilai berdasarkan ulasan yang masuk. HARRIS Hotel didirikan pada tahun 2001 di bawah Tauzia Hotel Management. Saat ini, HARRIS Hotel telah memiliki lebih dari 20 hotel yang tersebar di beberapa kota besar di Indonesia. Dan untuk wilayah Jawa Timur, HARRIS Hotel & Conventions Malang menjadi yang pertama berdiri pada tahun 2013. Berikut ini daftar lokasi Harris Hotel di Indonesia (sumber: https://mytauziaprivilege.com/collections)

1. HARRIS Resort Waterfront Batam
2. HARRIS Hotel Batam Center – Batam
3. HARRIS Hotel Tebet – Jakarta

4. HARRIS Hotel & Conventions Kelapa Gading – Jakarta
5. HARRIS Suites FX Sudirman – Jakarta
6. HARRIS Hotel & Conventions Bekasi
7. HARRIS Hotel Sentul City – Bogor
8. HARRIS Hotel & Conventions Festival CityLink – Bandung
9. HARRIS Hotel & Convention Ciumbuleuit – Bandung
10. HARRIS Hotel & Conventions Malang
11. HARRIS Hotel Tuban – Bali
12. HARRIS Hotel & Residences Riverview Kuta – Bali
13. HARRIS Hotel Seminyak – Bali
14. HARRIS Raya Kuta – Bali
15. HARRIS Denpasar – Bali
16. HARRIS Hotel & Residences Sunset Road – Bali
17. HARRIS Hotel Pontianak – Kalimantan
18. HARRIS Hotel Samarinda - Kalimantan

2.3 Dasar Teori

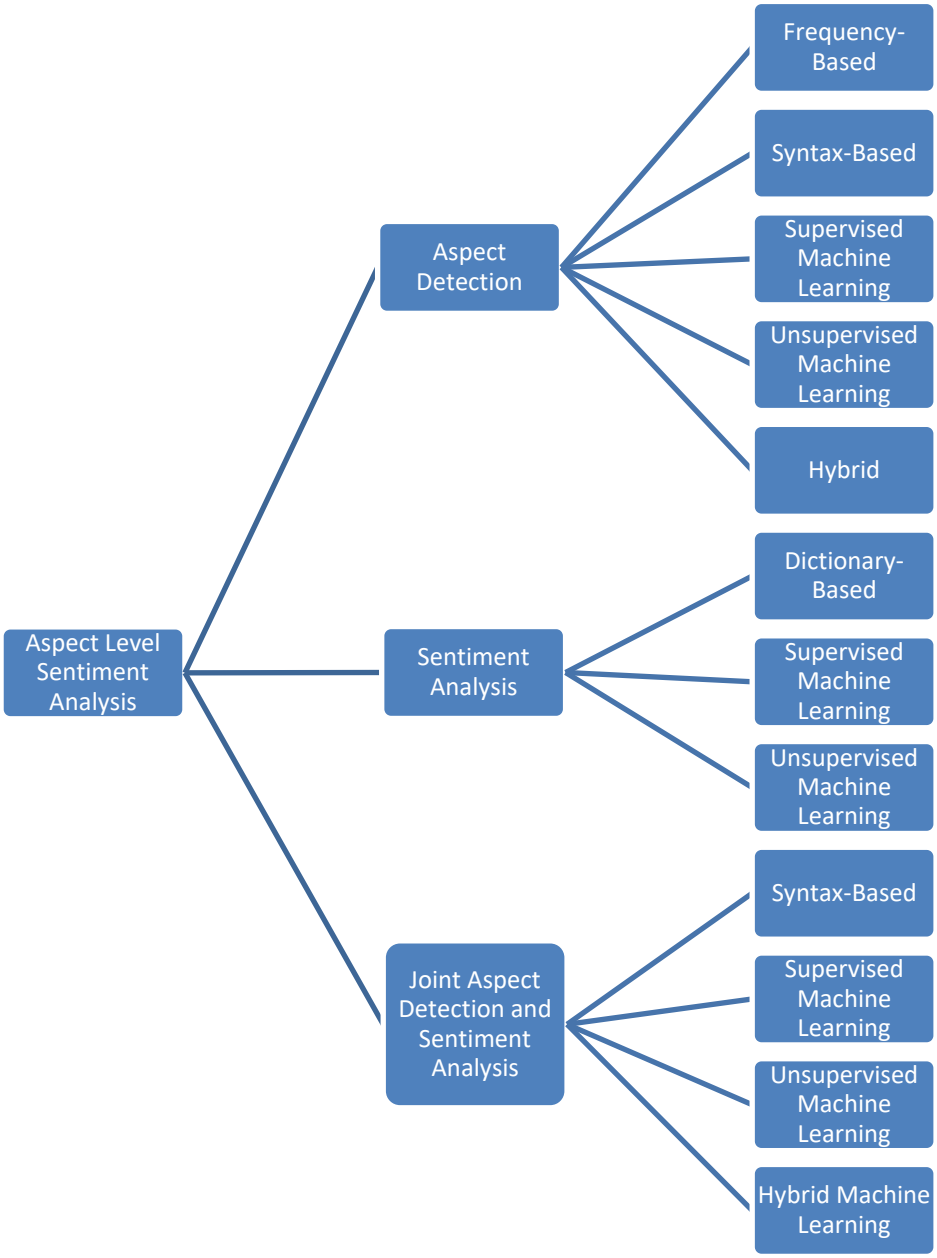
2.3.1 Analisis Sentimen Tingkat Aspek

Seseorang dapat mengekspresikan pendapatnya dalam berbagai tingkatan. Mulai dari ulasan secara umum hingga detail tentang aspek-aspek pada objek ulasan. Dari segi ruang lingkup untuk analisis sentimen ini, dapat terbagi ke dalam tiga tingkatan, yaitu tingkat dokumen, tingkat kalimat, dan tingkat aspek.

Untuk analisis sentimen tingkat dokumen, sentimen diambil berdasarkan keseluruhan bacaan yang ada di dokumen. Kemudian, untuk analisis sentimen tingkat kalimat terdapat dua tugas, yang pertama megidentifikasi apakah kalimat tersebut termasuk subjektif atau objektif dan yang kedua adalah mengklasifikasi kalimat yang subjektif apakah positif, negatif, atau netral (Westerski, A., 2007). Terakhir, tingkatan yang akan diteliti dalam penelitian ini, yaitu analisis sentimen tingkat aspek. Pada tingkatan ini, bukan hanya menentukan bagaimana sentimennya, tetapi juga pada aspek apa sentimen tersebut ditujukan.

Untuk aspek itu sendiri, ada yang menggunakan aspek yang sudah ditentukan sebelumnya dan ada yang dicari dengan menggunakan metode tertentu. Secara umum, proses analisis pada tingkat aspek memiliki tiga tahap, yaitu identifikasi, klasifikasi, dan agregasi (Schouten dan Frasincar, 2015). Dari survei yang dilakukan oleh Kim Schouten dan Flavius Frasincar disampaikan bahwa untuk pendekatan yang biasa dilakukan, terdapat tiga kategori, mulai dari metode yang fokus kepada identifikasi aspek, metode yang fokus kepada analisis sentimen, dan metode yang menggabungkan kedua tahapan tersebut. Metode yang fokus kepada identifikasi

aspek ada yang berdasarkan frekuensi, berdasarkan sintaks, *machine learning* terbimbing, *machine learning* tidak terbimbing, dan *hybrid* yaitu penggabungan dua metode atau lebih. Untuk metode yang fokus kepada analisis sentimen ada yang berdasarkan kamus, *machine learning* terbimbing, dan *machine learning* tidak terbimbing. Untuk metode yang fokus pada penggabungan keduanya ada berdasarkan sintaks, *machine learning* terbimbing, *machine learning* tidak terbimbing, dan *hybrid* yaitu penggabungan dua metode atau lebih. Gambar 2.1 ini dikutip dari survei yang dilakukan oleh Kim Schouten dan Flavius Frasincar dalam artikel ilmiah mereka yang berjudul "*Survey on Aspect-Level Sentiment Analysis*".



Gambar 2.1 Taksonomi untuk pedekatan pada analisis sentimen tingkat aspek

Sumber: Schouten dan Frasincar (2015)

2.3.2 Preprocessing

Merupakan tahapan yang dilakukan untuk mengolah data awal agar memberikan hasil yang baik nantinya saat diproses dengan metode yang sudah dipilih. Pada tahap ini, terdapat beberapa tahapan di dalamnya, yaitu:

2.3.2.1 Case Folding

Tahapan untuk mengubah semua huruf menjadi huruf kapital semua atau huruf kecil semua. Biasanya adalah menjadikan huruf kecil semua.

2.3.2.2 Filtering

Tahapan untuk menyaring kata-kata yang sering digunakan namun tidak memiliki makna yang penting (*stopwards removal).* Selain itu juga menghilangkan karakter yang diluar dari ketentuan, misal dari huruf a sampai z, selain dari itu dihapus.

2.3.2.3 Tokenizing

Tahapan untuk memisah setiap kata atau token dalam suatu rangkaian kata atau *input*. Untuk memisahkankannya dapat menggunakan spasi.

2.3.2.4 Part-of-Speech Tagging (POS Tagging)

Merupakan tahapan untuk menandai suatu kata atau frasa termasuk ke dalam *tag* atau kelas kata apa. Untuk melakukannya disini akan memanfaatkan antarmuka pemrograman aplikasi (API) yang disediakan oleh kateglo. Dengan cara melakukan akses dengan URL 'http://kateglo.com/api.php?format=json&phrase={kata}'. Dari respon yang didapat nantinya akan diambil nilai pada *key* 'lex_class_name' yang berisi nama kelas untuk kata tersebut.

2.3.3 Klasterisasi

Klasterisasi merupakan metode untuk megelompokkan sekumpulan data yang memiliki kemiripan. Dalam penelitian analisis sentimen tingkat aspek ini, diasumsikan kalimat-kalimat yang mirip akan memiliki aspek yang sama sehingga degan klasterisasi ini, akan didapat kelompok-kelompok kalimat yang mirip, artinya di dalam kelompok tersebut ada suatu aspek yang dibicarakan. Salah satu metode klasterisasi yang banyak diterapkan adalah K-Means. Namun untuk kasus dalam penelitian ini akan digunakan K-Means yang telah dimodifikasi yang disebut K-Modes.

2.3.3.1 K-Modes

K-Means (MacQueen, 1967) merupaka algoritme *unsupervised* yang sederhana yang hasilnya cukup baik untuk permasalahan klasterisasi yang umum. K-Means akan bekerja pada atribut atau fitur dengan nilai numerik namun tidak pada atribut dengan tipe kategorikal (Huang, 1998). Pada K-Modes yang berbeda adalah dari segi pengukuran ketidaksamaan (*dissimilarity measures)* dan penentuan *centroid* baru.

Secara umum algoritme K-Modes memiliki tahapan-tahapan dari penentuan jumlah klaster dengan memasukkan *centroid* awal, kemudian setiap objek dimasukkan ke dalam kelompok terdekat jaraknya dengan *centroid.* Jika semua objek sudah memiliki kelompok masing-masing, lakukan perhitungan untuk *centroid* baru dengan mengambil modus untuk setiap atribut pada setiap anggota kelompok. Ulangi pengelompokan hingga hasil pengelompokan tidak berubah atau nilai *centroid* tidak memiliki perubahan besar.

Untuk mengukur jarak dalam K-Modes menggunakan *dissimilarity measures* seperti yang dijelaskan pada Persamaan 2.1 dan Persamaan 2.2.

$$d1(X,Y) = \sum_{j=1}^{m} \delta\,(x_j, y_j) \tag{2.1}$$

Dimana

$$\delta(x_j, y_j) = \begin{cases} 0 & (\mathrm{x_j} = \mathrm{y_j}) \\ 1 & (\mathrm{x_j} \neq \mathrm{y_j}) \end{cases} \tag{2.2}$$

2.3.4 Klasifikasi

Klasifikasi adalah proses untuk memberi kelas pada setiap masukan secara terbimbing sesuai aturan atau kaidah yang dibuat. Bagaimana aturan atau proses pembelajara yang diterapkan tergantung dengan algoritme masing-masing. Dalam bidang jaringan saraf tiruan, terdapat metode-metode yang meniru cara kerja sistem saraf di otak sehingga memiliki keuntungan dalam pembelajaran yang berlanjut.

2.3.4.1 Score representation

Merupakan model ekstraksi fitur yang diusulkan oleh Farhadloo dan Rolland (2013) yang dalam penelitian mereka dapat meningkatkan akurasi cukup signifikan. $[S^+, S^0, S^-]^T$ adalah model ekstraksi fitur tersebut, dimana

$$S^+ = \sum_{i \in x} w_i s_i{}^+ \tag{2.3}$$
$$S^0 = \sum_{i \in x} w_i s_i{}^0$$
$$S^- = \sum_{i \in x} w_i s_i{}^-$$

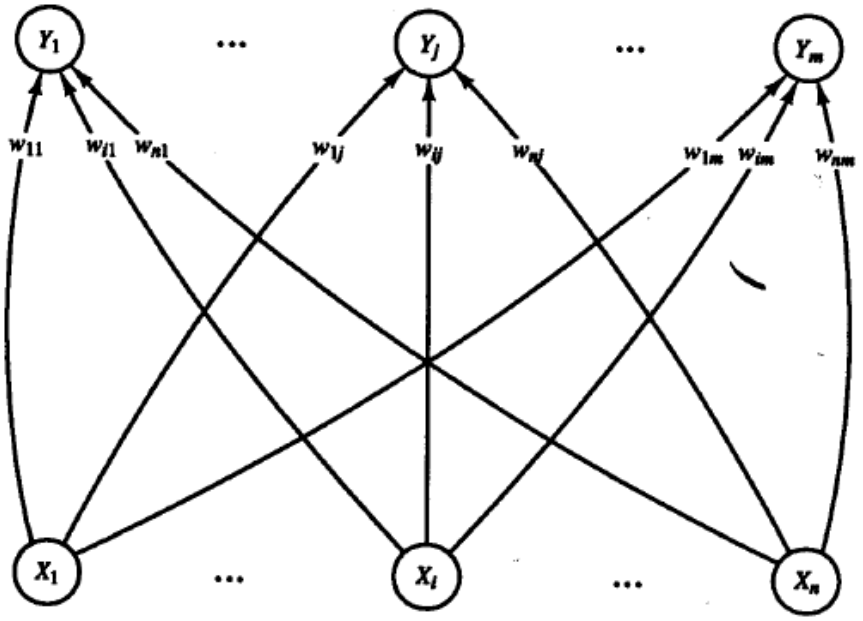
Satu kalimat terdiri dari tiga komponen yaitu nilai positif (S^+), nilai netral (S^0), dan nilai negatif (S^-). Nilai tersebut didapat dari penjumlahan hasil perkalian bobot setiap kata (w) dengan *score representation* masing-masing kata. Untup bobot (w) didapat dengan mengambil jumlah kemunculan kata.

$$s_i{}^+ = \frac{f_i^+}{f_i^+ + fi^0 + f_i^-}, \tag{2.4}$$
$$s_i{}^0 = \frac{f_i^0}{f_i^+ + fi^0 + f_i^-},$$
$$s_i{}^- = \frac{f_i^-}{f_i^+ + fi^0 + f_i^-}$$

Untuk *score representation* setiap nilai positif, netral, dan negatif, didapat dengan membagi frekuensi suatu kata pada kelasnya (positif, netral, atau negatif sesuai dengan pelabelan manual) dengan total frekuensi kata tersebut.

2.3.4.2 *Learning Vector Quantization 2* (LVQ2)

LVQ merupakan algoritme pembelajaran kompetitif yang disebut sebagai SOM versi *supervised* atau terbimbing. LVQ adalah metode klasifikasi dimana setiap unit *output* merepresentasikan sebuah kelas. Algoritme ini diusulkan oleh Kohonen pada tahun 1986. Kemudian muncul varian-varian LVQ1, LVQ2, LVQ2.1, LVQ3, Gambar 2.2 adalah model jaringan dasar dari LVQ.



Gambar 2.2 Model jaringan dasar LVQ

Sumber: Fausett (1994)

X = merupakan vector dari data latih

T = kategori atau kelas target (sesuai pelabelan manual) untuk setiap x

w_j = Vektor bobot untuk *output*

C_j = kategori atau kelas yang hasil pembelajaran dengan melihat kepada *output*

$||x - w_j||$ = Jarak Euclidean antara *input* dan bobot

Langkah 0.

Inisialisasi vektor awal untuk bobot awal dan inisialisasi *learning rate*.

Langkah 1.

Selama kondisi untuk berhenti belum terpenuhi, lakukan langkah 2-6

Langkah 2.

Untuk setiap vektor *input*, lakukan langkah 3-4

Langkah 3.

Cari nilai $||x\text{-}w_j||$ minimum sehingga diambil *output*

Langkah 4.

Perbaharui bobot w_j dengan ketentuan (2.5)

Jika $T = C_j$, maka

$$w_j(baru) = w_j(lama) + \alpha[x - w_j(lama)]$$

Jika $T \neq C_j$, maka

$$w_j(baru) = w_j(lama) - \alpha[x - w_j(lama)]$$

Langkah 5.

Pegurangan *learning rate*.

$\alpha = \alpha * Dec\alpha$ (2.6)

Langkah 6.

Tes apakah kondisi untuk berhenti terpenuhi. Kondisi untuk berhenti ini bisa dengan batas jumlah iterasi atau nilai *learning rate* sudah terlalu kecil.

Untuk versi LVQ 2 terdapat tambahan yaitu,

- Terdapat *winning unit* dan *runner-up* yang merupakan kelas yang berbeda
- Target kelas dari *input* sama dengan *runner-up*
- Jarak *input* dengan bobot untuk kelas *winning unit* hampir sama dengan bobot untuk kelas *runner-up*. Hal ini bisa diketahui jika memenuhi kedua syarat berikut,

$$\frac{d_C}{d_r} > 1 - \epsilon \quad (2.7)$$

dan

$$\frac{d_r}{d_C} < 1 + \epsilon$$

- Jika kondisi diatas terpenuhi maka, bobot akan diperbaharui dengan ketentuan,

$$w_{win}(baru) = w_{win}(lama) - \alpha[x - w_{win}(lama)]$$

$$w_{runnerup}(baru) = w_{runnerpup}(lama) + \alpha[x - w_{runnerup}(lama)]$$

2.3.5 Pengujian

2.3.5.1 Silhoutte Coefficient

Untuk melakukan pengujian terhadap algoritme klasterisasi, akan digunakan *Silhoutte Coefficient* (SC) dengan rumus sebagai berikut

(2.8)

$$SC = \frac{1}{N}\sum_{i=1}^{N} s(i)$$

(2.9)

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\ b(i)\}}$$

$s(i)$ = Nilai *silhouette* untuk objek i

$b(i)$ = Nilai minimum dari rata-rata jarak objek i dengan seluruh anggota dari setiap klaster diluar klaster tempat i berada

$a(i)$ = Nilai rata-rata jarak objek i dengan seluruh anggota lain di dalam klaster tempat i berada

Nilai $s(i)$ akan bernilai dalam rentang dari -1 sampai 1. Nilai $s(i)$ menunjukkan seberapa tepat objek tersebut berada dalam klasternya. Untuk contoh manualisasinya, pertama hitung nilai $a(i) - b(i)$ setiap data dengan mengelompokkannya per klaster. Tabel 4.21 menunjukkan contoh nilai $a(i) - b(i)$ pada satu data pada setiap klaster untuk keperluan perhitungan nilai *silhoutte*. Tanda (a) menunjukkan tempat klaster objek tersebut berada. Nila (b) diambil dari nilai minimum antar klaster di luar objek.

2.3.5.2 F1-Score

Untuk mengetahui kualitas dari metode yang diajukan maka diperlukan adanya sistem pengujian yang sesuai sehingga bisa terlihat bagaimana kinerja yang diberikan. Pada penelitian ini akan digunakan *precision, recall,* dan *f1-score* untuk mengevaluasi sistem.

Precision merujuk kepada seberapa akurat hasil klasifikasi dari keseluruhan haisl yang diperoleh. Sedangkan *recall* merujuk kepada seberapa akurat data yang benar diberikan berdasar basis data. Untuk *f1-score* adalah *harmonic mean* untuk *precision* dan *recall.*

(2.10)

Untuk menghitung *precision*, *True Positive / (True Positive + False Positive).*

Untuk menghitung *recall, True Positive / (True Positive + False Negative).*

Untuk *f1-score*, 2 × *Precision × Recall / (Precision + Recall).*