Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau State Islamic University of Sultan Syarif Kasim Riau
UIN SUSKA RIAU

BAB II
LANDASAN TEORI

2.1 Forum Diskusi

Objek dalam penelitian ini difokuskan pada beberapa forum diskusi online ataupun website yang menyediakan fitur diskusi bagi partisipan. Lebih khususnya penelitian ini ditujukan terhadap pengelolaan dan analisis komentar partisipan yang terlibat di dalam forum. Forum diskusi online memungkinkan seluruh partisipan dapat mengemukakan pendapat tanpa adanya batasan maupun interupsi dari partisipan lain, sehingga berpeluang meluasnya pembahasan dan menyebabkan meningkatnya jumlah komentar partisipan.

2.1.1 Pengertian Forum diskusi

Forum diskusi merupakan sebuah wadah yang dibuat sebagai tempat berbagi, berkomuniksi dan tempat untuk membahas suatu permasalah yang terjadi. Forum diskusi *online* merupakan forum diskusi yang diakses secara *online* (Wijayati, 2013).

2.1.2 Kolom omentar

Kolom komentar menyediakan *feedback* pembaca atau keterlibatan pembaca mengenai sebuah topik yang dibahas dalam sebuah forum dikusi. Komentar terbaru diasumsikan sebagai *query* atau kata kunci untuk menentukan tingkat kemiripan dengan komentar-komentar sebelumnya. Menurut Meishan dkk (2008), pada umumnya ada tiga keterhubungan antar komentar yang saling berhubungan dengan yang lainnya, yaitu :

> (1) *Topic Relation*, yaitu dua komentar yang berhubungan berdasarkan topik, jika keduanya membahas mengenai topik yang sama, sering dibuktikan dengan berbagi kata-kata umum terkait dengan topik. (2) *Quotation Relation*, dua dokumen yang saling berhubungan melalui

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau

kutipan jika satu komentar mengutip beberapa segmen teks dari komentar lain. Mengutip segmen teks meruapakan sebuah indikasi yang kuat bahwa komentar terbaru menanggapi komentar yang dikutip atau terlibat dalam diskusi. (3) *Mention Relation,* jika nama dari partisipan pada komentar sebelumnya dilampirkan pada komentar selanjutnya atau komentar terbaru, dua komentar ini terhubung berdasarkan *Mention.*

2.2 *Text Mining*

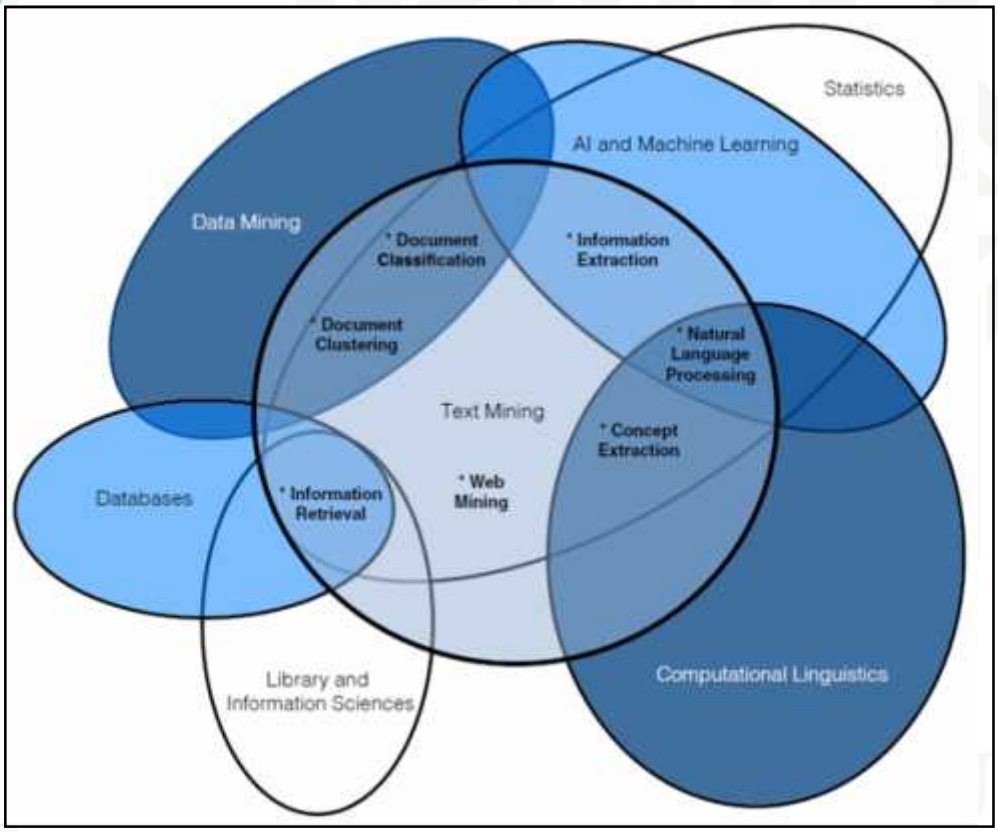
Text Mining atau *text analytics* adalah istilah yang mendeskripsikan subuah teknologi yang mampu menganalisis data teks semi-terstruktur maupun tidak terstruktur, hal inilah yang membedakannya dengan *data mining* dimana *data mining* mengolah daya yang sifatnya terstruktur. Pada dasarnya *text mining* merupakan bidang interdisiplin yang mengacu pada perolehan informasi (*information retrieval*), *data mining*, pembelajaran mesin (*mechine learning*), statistik, dan komputasi linguistik (Jiawei dkk, 2012 dikutip oleh Megawati, 2015). Tujuan utama dari *text mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Jadi, sumber data yang digunakan pada *text mining* adalah teks yang tidak terstruktur atau minimal semi terstruktur (Eldira, 2010). Aplikasi yang paling umum dilakukan *tekxt mining* saat ini misalnya penyaringan spam, analisis sentimen, mengukur preferensi pelanggan, meringkas dokumen, pengelompokan topik penelitian dan banyak lainnya. Menurut Miner dkk (2012) dikutip oleh penelitian Megawati (2015), pekerjaan *text mining* dikelompokkan menjadi 7 daerah praktek yang diilustrasikan pada gambar 2.1.

1. Pencarian dan perolehan informasi (*search and information retrieval*), yaitu penyimpanan dan penggalian dokumen teks misalnya dalam mesin pencarian (*search engine*) dan pencarian kata kunci (*keywords*).
2. Pengelompokan dokumen, yaitu penglompokan dan pengkategorian kata, istilah, paragraf, atau dokumen dengan mengunaka metode *cluster* (*clustering*) *data mining.*
3. Klasifikasi dokumen, yaitu pengelompokan dan pengkategorian kata, istilah, paragraf atau dokumen dengan menggunakan metode klasifikasi

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau

(*classification*) *data mining* berdasasrkan model terlatih yang sudah memiliki label.

4. *Web mining*, yaitu penggalian informasi dari internet dengan skala fokus yang spesifik.
5. Ekstraksi informasi (*information extraction*), yaitu mengidentifikasi dan mengekstraksi informasi dari data yang sifatnya semi-terstruktur atau tidak terstruktur dan mengubahnya menjadi data yang terstruktur.
6. *Natural language processing* (NLP), yaitu pembuatan program yang memiliki kemampuan untuk memahami bahasa manusia.
7. Ekstraksi konsep, yaitu pengelompokan kata atau frase ke dalam kelompok yang mirip secara semantik.



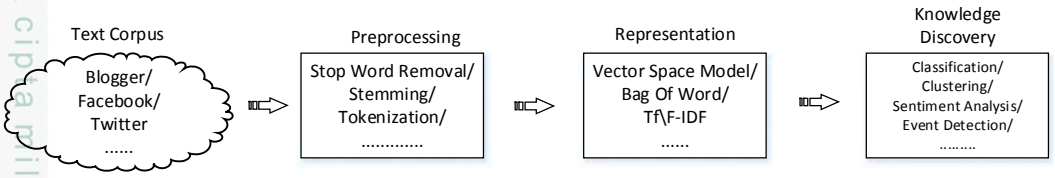
Gambar 2.1 Diagram Venn 6 bidang terkait dan 7 area praktek *text mining*

(Sumber: Miner dkk, 2012 dikutip oleh Megawati, 2015)

Pada penelitian ini kumpulan dokumen berupa kumpulan komentar yang terdapat dalam suatu diskusi berdasarkan topik bahasan. Komentar terbaru diasumsikan sebagai *query* atau kata kunci untuk menentukan kemiripan dengan

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau

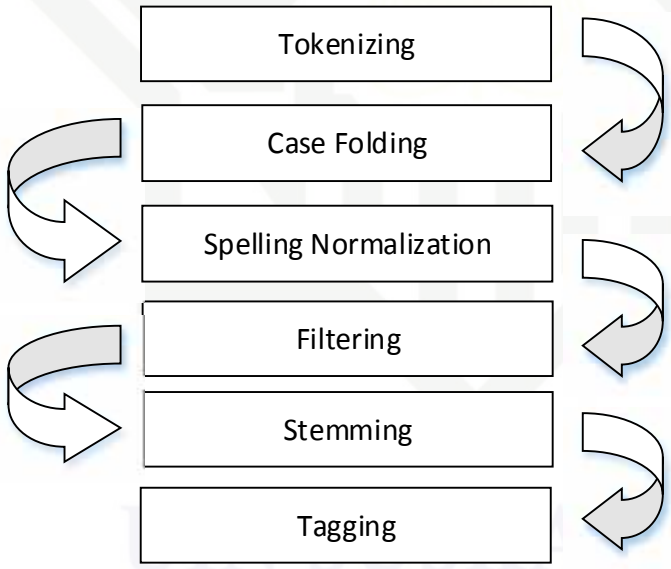
komentar sebelumnya. Proses-proses yang dilakukan dalam *text mining* secara umum adalah sebagai berikut:



Gambar 2.2 Kerangka Proses Analisis Teks Pada *Text Mining*
(Sumber : Zhai dkk , 2012 dikutip oleh Megawati, 2015)

2.2.2 Pra-proses (*Text Pre-processing*)

Data yang diinput terlebih dahulu akan melewati tahapan *pre-processing* untuk dapat dimengerti oleh sistem pengolahan *text mining* dengan baik. Tujuan utama tahapan *pre-processing* adalah untuk mendapatkan bentuk data siap oleh untuk diproses oleh sistem dari data awal berupa data tekstual. Gambar II.3 merupakan tahapan-tahapan *pre-processing*.

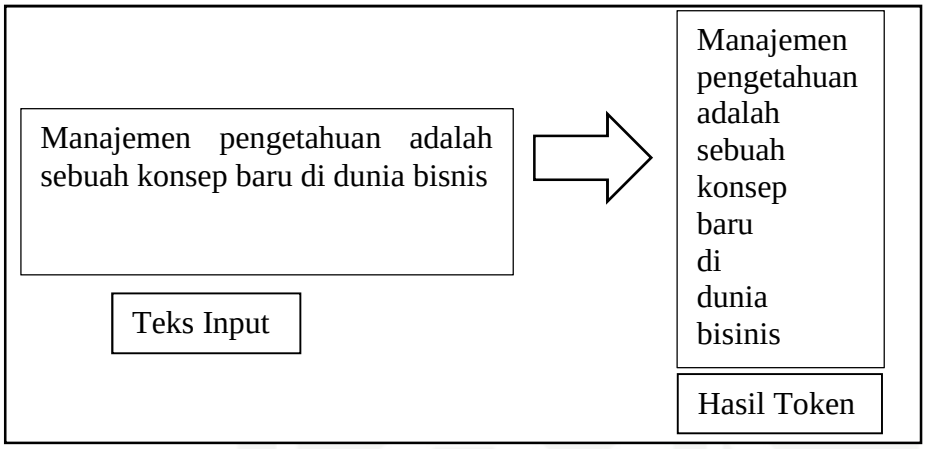


Gambar 2.3 Proses Text Mining
(Sumber : Megawati, 2015)

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau
UIN SUSKA RIAU

1. *Tokenizing*

Tokenizing adalah proses pemotongan string input berdasarkan kata yang menyusunnya.



Gambar 2.4 Proses Tokonisasi

(Sumber : Harlian, 2009)

2. *Case Folding*

Proses penyeragaman kosakata dengan mengubah huruf hasil tokenisasi dengan huruf kecil atau huruf besar pada keseluruhan kosakata.

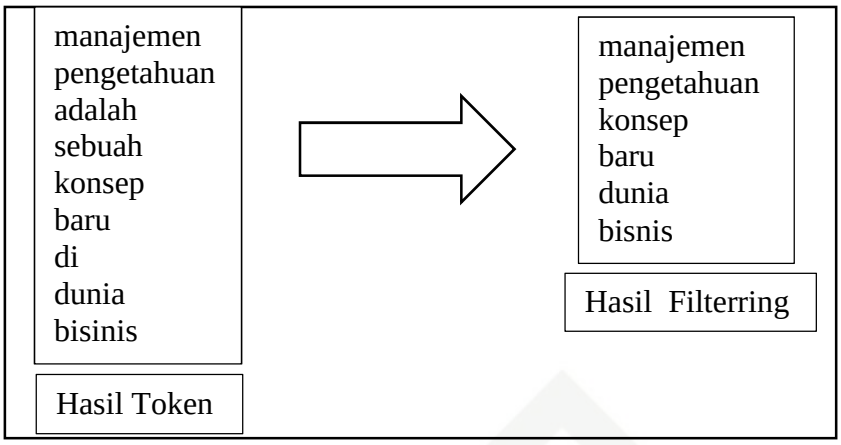
3. *Spelling Normalization*

Merupakan perbaikan dan subtitusi kata-kata yang salah eja ataupun disingkat dengan bentuk tertentu. Subtitusi kata dilakukan untuk menghindari jumlah perhitungan dimensi kata yang melebar. Perhitungan dimensi kata akan melebar jika kata yang salah eja atau disingkat tidak diubah karena kata tersebut sebenarnya memiliki kontribusi dalam merepresentasikan dokumen tetapi akan dianggap sebagai entitas yang berbeda pada proses penyusunan matriks.

4. Filtering

Filtering adalah tahap mengambil kata-kata penting dari hasil token. Biasanya tahap ini menggunakan algoritma *stop-list* (membuang kata-kata kurang penting) atau *word-list (*menyimpan kata penting*).*

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau

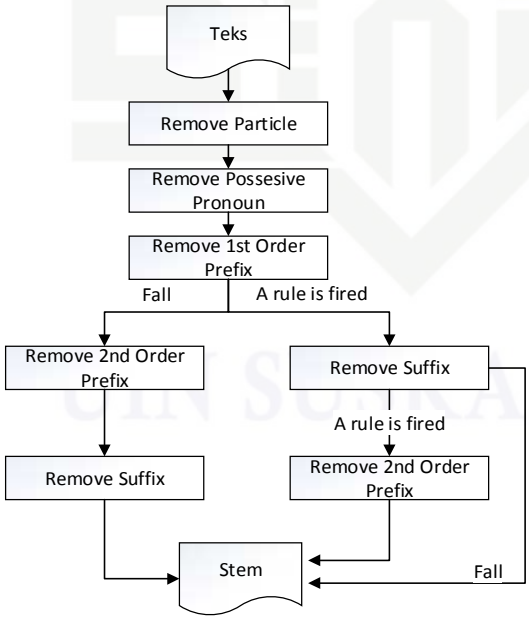


Gambar 2.5 Proses *Filtering*

(Sumber : Harlian, 2009)

5. *Stemming*

Stemming adalah tahap mencari *root*/ kata dasar dari setiap kata hasil dari proses *filtering.* Proses *stemming* dokumen teks berbahasa Indonesia menggunakan Algortima Porter membutuhkan waktu yang lebih singkat dibandingkan dengan *stemming* menggunakan Algoritma Nazief dan Adriani (Agusta, 2009). Proses *stemming* menggunakan Algoritma Porter dapat dilihat pada Gambar 2.6 berikut:



Gambar 2.6 Proses *Stemming* dengan Algoritma Porter

(Sumber : Agusta, 2009)

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau

Berikut adalah penjelasan mengenai alur dari algoritma Porter-Stemmer berdasarkan Gambar 2.6 di atas:

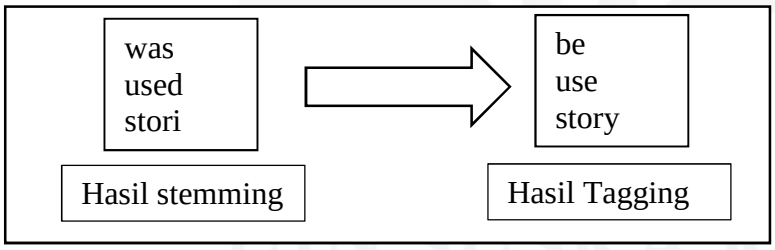
1. Menghapus partikel seperti: -kah, -lah, -tah
2. Mengapus kata ganti (*Possesive Pronoun*), seperti –ku, -mu, -nya
3. Mengapus awalan pertama. Jika tidak ditemukan, maka lanjut ke langkah 4a, dan jika ada maka lanjut ke langkah 4b.
4. a. Menghapus awalan kedua, dan dilanjutkan pada langkah ke 5a

   b. Menghapus akhiran, jika tidak ditemukan maka kata tersebut diasumsikan sebagai kata dasar (*root word*). Jika ditemukan maka lanjut ke langkah 5b.
5. a. Menghapus akhiran dan kata akhir diasumsikan sebagai kata dasar (*root word*).

   b. Menghapus awalan kedua dan kata akhir diasumsikan sebagai kata dasar (*root word*).

6. *Tagging*

Tagging adalah tahap mencari bentuk awal/ *root* dari kata lampau atau kata hasil *stemming*. Untuk dokumen berbahasa Indonesia proses *tagging* tidak diterapkan, karena Bahasa Indonesia tidak memiliki bentuk lampau.



Gambar 2.7 Proses *Tagging*

(Sumber : Harlian, 2009)

2.2.3 Penyusunan Vektor (*Representaion*)

1. TF/ IDF

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau
UIN SUSKA RIAU

Tahap analyzing adalah tahap penetuan keterkaitan kata-kata antar dokumen. Untuk melakukan analisa pada tahap *analyzing* dapat digunakan Algoritma TF/ IDF (*Term Frequency-Inverse Documents Frequency*) dan Algoritma *Vector Space Model.*

Algoritma TF/ IDF digunakan untuk menghitung bobot (W) masing-masing dokumen terhadap kata kunci dengan persamaan (Manning dkk, 2009) :

$$W_{dt} = tf_{dt} * IDF_t \quad (2.1)$$

$$IDF_t = \log(\frac{|D|}{DF_t}) \quad (2.2)$$

Dimana :

d = dokumen ke-d

t = kata ke-t dari kata kunci

W = bobot dokumen ke-d terhadap kata ke-t

tf = banyaknya kata yang dicari pada sebuah dokumen

IDF = Inversed Documents Frequency

D = total dokumen

df = banyaknya dokumen yang mengandung kata yang dicari

Setelah bobot (W) masing-masing dokumen diketahui, maka dilakukan proses *sorting/* pengurutan dimana semakin besar nilai W, semakin besar tingkat kemiripan dokumen tersebut terhadap kata kunci, demikian sebaliknya.

2. Normalisasi Bobot

Normalisasi bobot dokumen bertujuan untuk menangani melebarnya dimensi matriks dokumen dikarenakan pengulangan kata-kata yang sama dalam satu dokumen yang sama. Normalisasi berusaha menghilangkan efek panjang dokumen (Manning dkk, 2009).

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

$$w'_{i,j} = \frac{w_{i,j}}{\sqrt{w_{i,j}^2}} \quad (2.3)$$

$w'_{i,j}$ = bobot kata ke-i yang dinormalisasi

$w_{i,j}$ = bobot kata ke-i sebelum normalisasi

2.2.4 Ekstraksi Informasi pada *Text Mining*

Tahap akhir penggalian informasi pada *text mining* yaitu ekstraksi ilmu pengetahuan (*knowledge discovery*), dimana terdapat beberapa jenis kategori utama yang bisa dilakukan sebagai berikut (Miner dkk, 2012 dikutip oleh Megawati, 2015).

1. Klasifikasi/ prediksi,

Klasifikasi adalah bentuk analisis data yang mengekstrak model untuk menggambarkan kelas data (Jiawei dkk, 2012 dikutip oleh Megawati, 2015). Model yang dibangun meliputi pengklasifikasian dan prediksi kategori label kelas. Klasifikasi data mempunyai dua tahapan proses, yaitu tahap pembelajaran (*learning step*) dimana model klasifikasi dibangun berdasarkan label yang sudah diketahui sebelumnya dan tahapan klasifikasi (*classification step*) dimana model digunakan untuk memprediksi label kelas dari data yang diberikan (Miner dkk, 2012 dikutip oleh Megawati, 2015). Klasifikasi memiliki berbagai aplikasi, termasuk deteksi penipuan, penargetan marketing, prediksi kinerja, manufaktur, diagnosis medis, dan banyak lainnya. Sebagai contoh, kita dapat membangun sebuah model klasifikasi untuk mengkategorikan apakah suatu aplikan pinjaman bank termasuk aman atau beresiko. karena pada awal pembangunan model label kelas dari data telah diketahui, klasifikasi juga disebut sebagai metode *supervised learning.*

2. Pengelompokan (*Clustering*)

Tidak seperti klasifikasi, pada model *clustering* pengelompokan data dilakukan demga menggunakan algortma yang sudah ditentukan dan selanjutnya data akan diproses oleh algoritma untuk dikelompokkan menurut karakteristik

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

alaminya. Tidak ada unsur pembimbingan (dengan pemberian label kelas), melainkan algoritma akan berjalan dengan sendirinya untuk mengelompokkan data tersebut. Data yang lebih dekat (mirip) dengan data lain akan berkelompok dalam satu *cluster*, sedangkan data yang lebih jauh (berbeda) dari data yang lain akan berpisah dalam kelompok yang berbeda. Untuk masalah pengelompokkan data berdasarkan kemiripan/ ketidakmiripan antar data tanpa ada label kelas yang diketahui sebelumnya disebut dengan pembelajaran tidak terbimbing atau *unsupervised learning.* Dalam konteks yang lain, pembelajaran tidak terbimbing disebut juga pengelompokan atau *clustering*. Menurut struktur, *clustering* terbagi menjadi dua, yaitu *hierarki* dan *partisi*. Dalam pengelompokan berbasis hierarki, satu data tunggal bisa dianggap sebuah *cluster*, dua atau lebih *cluster* kecil dapat bergabung menjadi sebuah *cluster* besar, begitu seterusnya hingga semua data dapat bergabung menjadi sebuah *cluster*. Di sisi lain, pengelompokan berbasis partisi membagi set data ke dalam sejumlah *cluster* yang tidak bertumpang-tindih antara satu *cluster* dengan *cluster* yang lain, artinya setiap data hanya menjadi anggota satu *cluster* saja.

3. Asosiasi

Asosiasi merupakan proses pencarian hubungan antar elamen data. Dalam dunia industri retail, analisis asosiasi biasanya disebut *market Basket Analysis* (Miner dkk, 2012 dikutip oleh Megawati, 2015). Asosiasi tersebut dihitung berdasarkan ukuran *Support* (presentase dokumen yang memuat seluruh konsep suatu produk A dan B) dan *confidence* (presentase dokumen yang memuat seluruh konsep produk B yang berada dalam subset yang sama dengan dokumen yang memuat seluruh konsep produk A).

4. Analisis Tren

Tujuan dari analisis tren yaitu untuk mencari perubahan suatu objek atau kejadian oleh waktu (Miner dkk, 2012 dikutip oleh Megawati, 2015). Salah satu aplikasi analisis tren yaitu kegiatan identifikasi evolusi topik penelitian pada artikel akademis.

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

2.3 *Clustering*

*Cluster*ing merupakan salah satu teknik data mining yang bertujuan untuk mengidentifikasi sekelompok obyek yang mempunyai kemiripan karakteristik tertentu yang dapat dipisahkan dengan kelompok obyek lainnya, sehingga obyek yang berada dalam kelompok yang sama relatif lebih homogen daripada obyek yang berada pada kelompok yang berbeda. Tujuan dari pengelompokan sekumpulan data obyek ke dalam beberapa kelompok yang mempunyai karakteristik tertentu dan dapat dibedakan satu sama lainnya adalah untuk analisis dan interpretasi lebih lanjut sesuai dengan tujuan penelitian yang dilakukan.

2.3.1 *Hierarchical Clustering* (Metode Hirarki)

Tekinik hirarki adalah teknik *cluster*ing membentuk konstruksi hirarki atau berdasarkan tingkatan tertentu seperti struktur pohon (struktur perbandingan). Dengan demikian proses pengelompokkan dilakukan secara bertingkat atau bertahap. Hasil dari pengelompokkan ini dapat disajikan dalam bentuk dendogram. Metode-metode yang digunakan dalam bentuk teknik hirarki:

1. Agglomerative clustering

Metode ini dimulai dengan kenyataan bahwa setiap obyek membentuk *cluster*nya masing-masing. Kemudian dua obyek bergabung dengan jarak terdekat bergabung. Selanjutnya obyek ketiga bergabung dengan *cluster* yang ada atau bersama obyek lain dan membentuk *cluster* baru. Hal ini tetap memperhitungkan jarak kedekatan antar obyek. Proses akan berlanjut hingga akhirnya akan terbentuk satu *cluster* yang tediri dari keseluruhan obyek. Beberapa metode dalam *Agglomerative clustering* yaitu :

a. Single Linkage Hierarchical Method

Metode ini menggunakan prinsip yang diawali dengan mencari dua obyek terdekat dan keduanya membentuk *cluster* yang pertama. Pada langkah selanjutnya terdapat dua kemungkinan, obyek ketiga akan bergabung dengan *cluster* yang telah terbentuk atau dua obyek lainnya akan membentuk *cluster* baru.

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

b. Complete Linkage Hierarchical Method

Metode ini kebalikan dari pendekatan yang digunakan pada *single linkage.* Prinsip jarak yang digunakan adalah jarak terjauh antar obyek.

c. Average Linkage Hierarchical Method

Metode ini mengikuti prosedur yang sama dengan kedua metode sebelumnya. Prinsip ukuran jarak yang digunakan adalah rata-rata antar tiap pasangan obyek yang mungkin.

d. Centroid Linkage Hierarchical Method

Pada *centroid clustering* kemiripan dua obyek ditentukan dengan ukuran jarak antar *centroid*nya.

2. Divisive Clustering

Metode *divisive* berlawanan dengan metode *agglomerative.* Metode ini pertama-tama diawali dengan satu *cluster* besar yang mencakup semua observasi (obyek). Selanjutnya obyek yang mempunyai tidak kemiripan yang cukup besar akan dipisahkan sehingga membentuk *cluster* yang lebih kecil. Pemisahan ini dilanjutkan sehingga mencapai sejumlah *cluster* yang diinginkan.

2.3.2 *Partitioning Clustering/ Non-Hierarchical Clustering*

Metode ini dimulai dengan memilih sejumlah nilai *cluster* awal sesuai dengan jumlah yang diinginkan dan kemudia obyek digabungkan ke dalam *cluster* tersebut. Beberapa contoh metode *non-hierarchical* ini adalah, *Sequential Treshold Procedure, Parallel Treshold Procedure, Optimizing.* Teknik partisi ini mencakup *K-Means Clustering* dan *Methods Based on the Trace.*

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau

2.4 Peng*cluster*an komentar Secara otomatis (*Automatic Comments Clustering*)

Setelah melewati *pre-processing* teks *mining*, maka akan dilajutkan pada tahap peng*cluster*an komentar. Pada tahap ini melibatkan algoritma *centroid lingkage hierarchical method* (CLHM) dan Metode *Hill-Climbing*.

2.4.1 *Centroid Linkage Hierarchical Method* (CLHM)

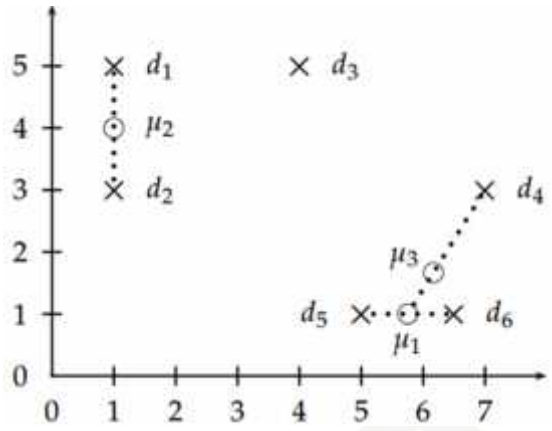
Centroid Linkage Hierarchical Method adalah proses peng*cluster*an yang didasarkan pada jarak antar *centroid*nya (Manning dkk, 2009). Metode ini baik untuk kasus *clustering* dengan normal data set distribution. Akan tetapi metode ini tidak cocok untuk data yang mengandung *outlier*. Langkah-langkah dalam proses peng*cluster*an dengan metode ini diantaranya adalah:

1. Diasumsikan setiap data dianggap sebagai *cluster*. Jika n=jumlah data dan c=jumlah *cluster*, maka c=n.
2. Menghitung jarak antar *cluster* danga jarak *cosine similarity* dan *euclidean.*
3. Mencari 2 *cluster* yang mempunyai jarak *centroid* antar *cluster* yang paling minimal dan digabungkan ke dalam *cluster* baru (sehingga c= c-1).
4. Kembali ke langkah 3, dan ulangi sampai dicapai *cluster* yang diinginkan.
5. Penghitungan jarak antar obyek, maupun antar *cluster*nya dilakukan dengan jarak *euclidean* (Persamaan 2.4) dan *cosine similarity* (Persamaan 2.5).

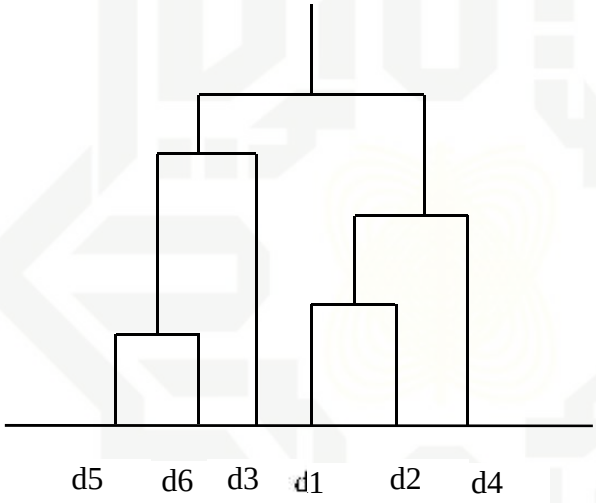
Gambar 2.8 berikut merupakan ilustrasi proses Algoritma CLHM pada proses *cluster*. Gambar 2.9 merupakan *dendogram* yang menggambarkan hasil dari proses *cluster*.

UIN SUSKA RIAU

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau
UIN SUSKA RIAU



Gambar 2.8 Ilustrasi Algoritma *Centroid Linkage Hierarchical method*
(Sumber : Manning dkk, 2009)



Gambar 2.9 *Dendogram* Algoritma *Centroid Linkage Hierarchical Method*

Berikut adalah dua metode pengukuran jarak antar dokumen/ komentar pada penelitian ini:

1. Jarak *Euclidean*

Jarak *euclidean* merupakan ukuran ketidakmiripan antar obyek yang diformulasikan oleh persamaan berikut:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \quad (2.4)$$

x, y ∈ X dan x_i, y_i adalah fitur ke-*i* dari x dan y, sedangkan n adalah jumlah fitur dalam vektor. Ukuran ketidakmiripan *euclidean* akan memberikan

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

d_0=0, maka jarak minimal yang mungkin di antara dua vektor data juga 0. Selain itu, jarak dari x ke y akan sama dengan y ke x, d(x,y) = d(y,x).

2. Cosine Similarity

Matriks *cosine similarity* merupakan ukuran kemiripan yang sering digunakan, misalnya untuk mengukur kemiripan dokumen *d* dan *q* (Prasetyo, 2014). Kemiripan yang diberikan adalah 1 jika dua vektor d dan q sama, dan bernilai 0 jika kedua vektor d dan q berbeda. Nilai jarak 1 menyatakan sudut yang dibentuk oleh vektor d dan q adalah 0^0, yang artinya vektor d dan q adalah sama (dalam hal jarak). Jika mendekati 0 berarti sudut yang dibentuk vektor d dan q adalah 90°, yang artinya tidak ada sesuatupun yang sama di antara vektor d dan q. Ukuran kemiripan *cosine similarity* diformulasikan dengan Persamaan 2.5 berikut (Prasetyo, 2014):

$$Cosin\ e \rightarrow sim\ d,q\ = \frac{\vec{d} * \vec{q}}{\vec{d} \times |\vec{q}|} = \frac{\sum_{i=1}^{t}(W_{id} \times W_{iq})}{\sum_{i=1}^{t} W_{id}^2 \times \sum_{i=1}^{t} W_{iq}^2} \quad (2.5)$$

Dimana :

$\vec{d}$ = vektor dokumen d

$\vec{q}$ = vektor dokumen q

W_{id} = Bobot kata ke-i pada dokumen d

W_{iq} = Bobot kata ke-i pada dokumen q

t = jumlah kata/ *term*

2.4.2 Analisa *Cluster*

Analisa *cluster* adalah suatu teknik analisa *multivariative* (banyak variabel) untuk mencari dan mengorganisir informasi tentang variabel tersebut sehingga secara relatif dapat dikelompokkan dalam bentuk yang homogen dalam sebuah *cluster* (Eldira, 2010). Analisa *cluster* bisa diperoleh dari kepadatan *cluster* yang dibentuk (*cluster density*). Kepadatan suatu *cluster* bisa ditentukan dengan *variance whitin cluster (Vw)* dan *variance between cluster* (Vb).

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau
UIN SUSKA RIAU

Varian tiap tahap pembentukan *cluster* bisa dihitung dengan Persamaan(2.6) *(*Martiana dkk,2010*)* :

$$Vc^2 = \frac{1}{n_c - 1} \sum_{i=1}^{n} (y_i - \bar{y}_c)^2 \quad (2.6)$$

Dimana:

Vc^2 = varian pada *cluster* c

c =1...k, dimana k=jumlah *cluster*,

n_c =jumlah data pada *cluster* c,

y_i =data ke-i pada suatu *cluster*,

y_c = rata-rata dari data pada suatu *cluster*.

Selanjutnya dari nilai varian di atas, maka dapat dihitung nilai dari *variance within cluster* (Vw) dengan Persamaan (2.7) *(*Martiana dkk,2010*)* :

$$Vw = \frac{1}{N - c} \sum_{i=1}^{c} (n_i - 1)\, V_i^2 \quad (2.7)$$

Dimana:

N= jumlah semua data,

$n_{i=}$jumlah data *cluster* i,

$V_{i=}$varian pada *cluster* i.

Dan nilai variance between cluster (Vb) dengan Persamaan(2.8) (Martiana dkk, 2010):

$$Vb = \frac{1}{c - 1} \sum_{i=1}^{c} n_i (\bar{y}_i - \bar{y})^2 \quad (2.8)$$

Dimana:

$\bar{y} = rata - rata\, dari\, \bar{y}_i$

Salah satu metode yang digunakan untuk menentukan *cluster* yang ideal adalah batasan varian, yaitu dengan menghitung kepadatan *cluster* berupa *variance within cluster* (Vw) dan *varian between cluster* (Vb). *Cluster* yang ideal

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

mempunyai Vw minimum yang mempresentasikan *internal homogenety* dan maksimum Vb yang menyatakan *external homogenety.*

$$V = \frac{Vw}{Vb} \tag{2.9}$$

Meskipun minimum Vw menunjukkan nilai *cluster* ideal, tetapi pada beberapa kasus kita tidak bisa menggunakannya secara langsung untuk mencapai *global optimum*. Jika kita paksakan, maka solusi yang dihasilkan akan jatuh pada *local optima.*

2.4.3 Identifikasi Pola Berdasarkan Pergerakan Varian

Identifikasi pola pergerakan varian merupakan metode untuk memperoleh *cluster* yang mencapai *global optimum,* yang mampu mengatasi masalah dari minimum V.

Posisi yang mungkin untuk menemukan *global optimum* pada pergerakan varian, dikelompokkan menjadi 2, yaitu:

1. Hill-climbing

Berdasarkan penilitian Eldira dkk (2010), metode *Hill-climbing* memiliki kemampuan yang lebih baik dari pada metode *valley-tracing* dalam mengenali pola varian *cluster* diantaranya adalah:

(1) Jumlah *cluster* yang terbentuk relatif lebih banyak karena sensitifitas tiap *cluster*. (2) Kecepatan *running* karena anggota tiap *cluster* sedikit sehingga proses pencarian dokumen lebih cepat. (3) Metode *Hill-climbing* sangat cocok untuk kata kunci yang panjang dan data yang banyak.

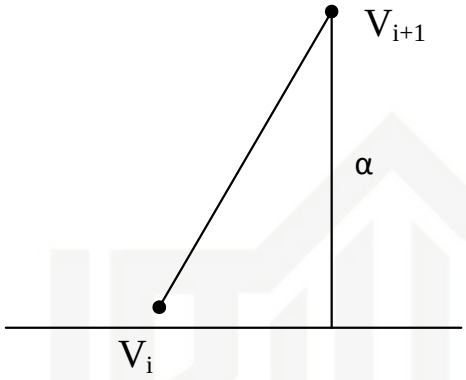
Pada *hill-climbing* didefinisikan bahwa kemungkinan mencapai *global optimum* terletak pada tahap ke-i, jika memenuhi Persamaan (2.10) berikut: (Eldira dkk, 2010):

$$V_{i+1} > a.V_i \tag{2.10}$$

Dimana a adalah nilai tinggi.

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau
UIN SUSKA RIAU

Nilai tinggi digunakan untuk menentukan seberapa mungkin metode ini mencapai *global optimum.* Nilai a yang bisa digunakan adalah 2, 3 dan 4. Persamaan (2.10) di atas, diperoleh berdasarkan analisa pergerakan varian pola *Hill-climbing* berikut:



Gambar 2.10 Pola Nilai Beda *Hill-climbing*
(Sumber : Eldira dkk, 2010)

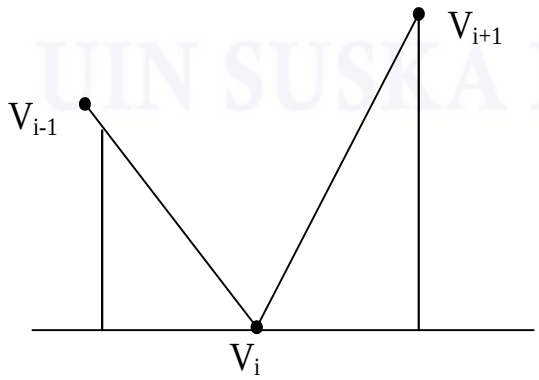
2. Valley-Tracing

Pada *Valley-Tracing* didefinisikan bahwa kemungkinan mencapai *global-optimum* terletak pada tahap ke-i, jika memenuhi persamaan berikut (Eldira dkk, 2010):

$$(v_{i-1} + v_i) \cap (v_{i+1} + v_i) \quad (2.11)$$

Dimana i=1...n dan n tahap terakhir pembentukan *cluster*.

Persamaan diatas, diperoleh berdasarkan analisa pergerakan varian pola *valley-tracing* berikut:



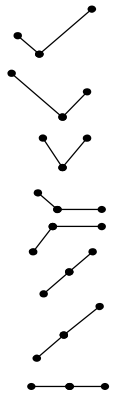
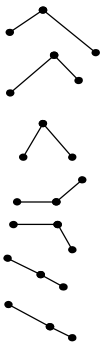
Gambar 2.11 Pola Nilai Beda *Valley-tracing*

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau
UIN SUSKA RIAU

(Sumber : Martiana dkk, 2010)

Tabel 2.1 berikut menunjukkan pola-pola *Hill-climbing* dan *valley-tracing* yang mungkin mencapai *global-optimum.*

Tabel 2.1 Kemungkinan pola *Hill-climbing* mencapai *global-optimum*

Pola	Mungkin	Pola	Mungkin
	✓ ✓ ✓ x x x x x		x x x ✓ x x x x

Selanjutnya, baik dengan pendekatan metode *valley-tracing* maupun *Hill-climbing* dilakukan identifikasi perbedaan nilai tinggi(∂) pada tiap tahap dengan Persamaan (2.12). Nilai ∂ digunakan untuk menghindari *local-optima*, dimana persamaan ini diperoleh dari maksimum ∂ yang dipenuhi pada Persamaan (2.12). Untuk membentuk *cluster* secara otomatis, yaitu *cluster* yang mencapai *global-optima*, digunakan nilai λ sebagai *treshold*, sehingga *cluster* secara otomatis terbentuk ketika memenuhi Persamaan (2.13) (Eldira dkk, 2010).

$$\partial = V_{i+1} - (V_i \square \alpha) \quad (2.12)$$

$$\text{Max}(\partial) \geq \lambda \quad (2.13)$$

2.4.4 Validitas *Cluster*

Berikut adalah 2 tahapan yang diterapkan dalam mengevaluasi *cluster* dengan mengukur kriteria tertentu sehingga diketahui validitas *cluster* yang terbentuk:

1. Akurasi *cluster*

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

Untuk mengetahui keakuratan dari sutau metode pembentukan *cluster* pada *hierarchical method,* baik menggunakan *valley-tracing* maupun *hill-climbing,* digunakan Persamaan (2.14), dengan nilai terdekat ke max (∂) adalah nilai kandidat max (∂) sebelumnya. Nilai phi (φ) yang lebih besar atau sama dengan 2 ($\varphi \geq 2$), menunjukkan *cluster* yang tebentuk merupakan *cluster* yang *well-separated* (terpisah dengan baik), (Eldira dkk, 2010).

$$\varphi = \frac{\text{Max}(\partial)}{nilai\ terdekat\ ke\ \text{Max}(\partial)} \tag{2.14}$$

Tahap terakhir adalah proses *sorting* (pengurutan) untuk mengurutkan komentar yang memiliki kemiripan tertinggi hingga terendah.

2. Validitas Eksternal

Validitas eksternal merupakan pengukuran validitas *cluster* dengan melibatkan informasi eksternal, dalam penelitian ini informasi eksternal tersebut adalah informasi kelas komentar yang memiliki kemiripan yang disajikan dalam bentuk mariks *confusion*. Matriks confusion berisi informasi mengenai dokumen yang relevan maupun yang tidak relevan dan dokumen yang ditemukan dan tidak ditemukan.

Tabel 2.2 Matriks *Confusion*

	Relevant	*Not Relevant*
Retrieved	A	B
Not Retrieved	C	D

1. a menunjukkan bahwa dokumen ter-*cluster* dan relevan dengan kelas sebenarnya.
2. b menunjukkan bahwa dokumen ter-*cluster* tapi tidak relevan dengan kelas sebenarnya.
3. c menunjukkan bahwa dokumen tidak ter-*cluster* tapi dokumen relevan dengan kelas sebenarnya.
4. d menunjukkan bahwa dokumen tidak ter-*cluster* dan tidak relevan dengan kelas sebenarnya.

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

Ada beberapa standar pengukuran yang digunakan dalam *clustering* dokumen, diantaranya adalah *recall, precision, F-measure.*

1. Recall

Recall adalah tingkat kebehasilan mengenali suatu *event* dari seluruh *event* yang seharusnya dikenali. Cara mencari nilai *recall* ini adalah dengan membagi jumlah dokumen yang ditemukan dan relevan dengan jumlah seluruh dokumen yang sebenarnya relevan. Metode *recall* dapat dirumuskan dengan Persamaan (2.15) berikut :

$$Recall = \frac{|\ Relevant \cap Retrieved\ |}{|\ Relevant\ |} \quad (2.15)$$

Keterangan :

Recall: Tingkat keberhasilan

{*Relevant*} : Kumpulan dokumen yang relevan

{*Retrieved*}: Kumpulan dokumen yang ditemukan

2. Precision

Precision adalah tingkat ketepatan hasil *cluster* terhadap suatu kejadian. Metode *precision* ini dapat dirumuskan seperti pada Persamaan (2.16) berikut:

$$Precision = \frac{Relevant \cap Retrieved}{Retrieved} \quad (2.16)$$

Keterangan:

Precision: Tingkat ketapatan

{*Relevant*} : Kumpulan dokumen yang relevan

{*Retrieved*}: Kumpulan dokumen yang ditemukan

3. F-Measure

UIN SUSKA RIAU
Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau
UIN SUSKA RIAU

F Measure merupakan gabungan antara *recall* dan *precision* yang didefinisikan dengan Persamaan (2.17) berikut :

$$F\ Measure \frac{2\ x\ Recall\ x\ Precision}{Recall + Precision} \quad (2.17)$$

© Hak cipta milik UIN Suska Riau State Islamic University of Sultan Syarif Kasim Riau

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
   a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
   b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.