BAB II

LANDASAN TEORI

2.1 Tinjauan Pustaka

Dalam penyusunan skripsi telah dilakukan tinjauan pustaka terhadap 7 penelitian terdahulu yang terangkum dalam bentuk *literature review* pada Tabel 2.1.

Tabel 2.1 *Literature Review*

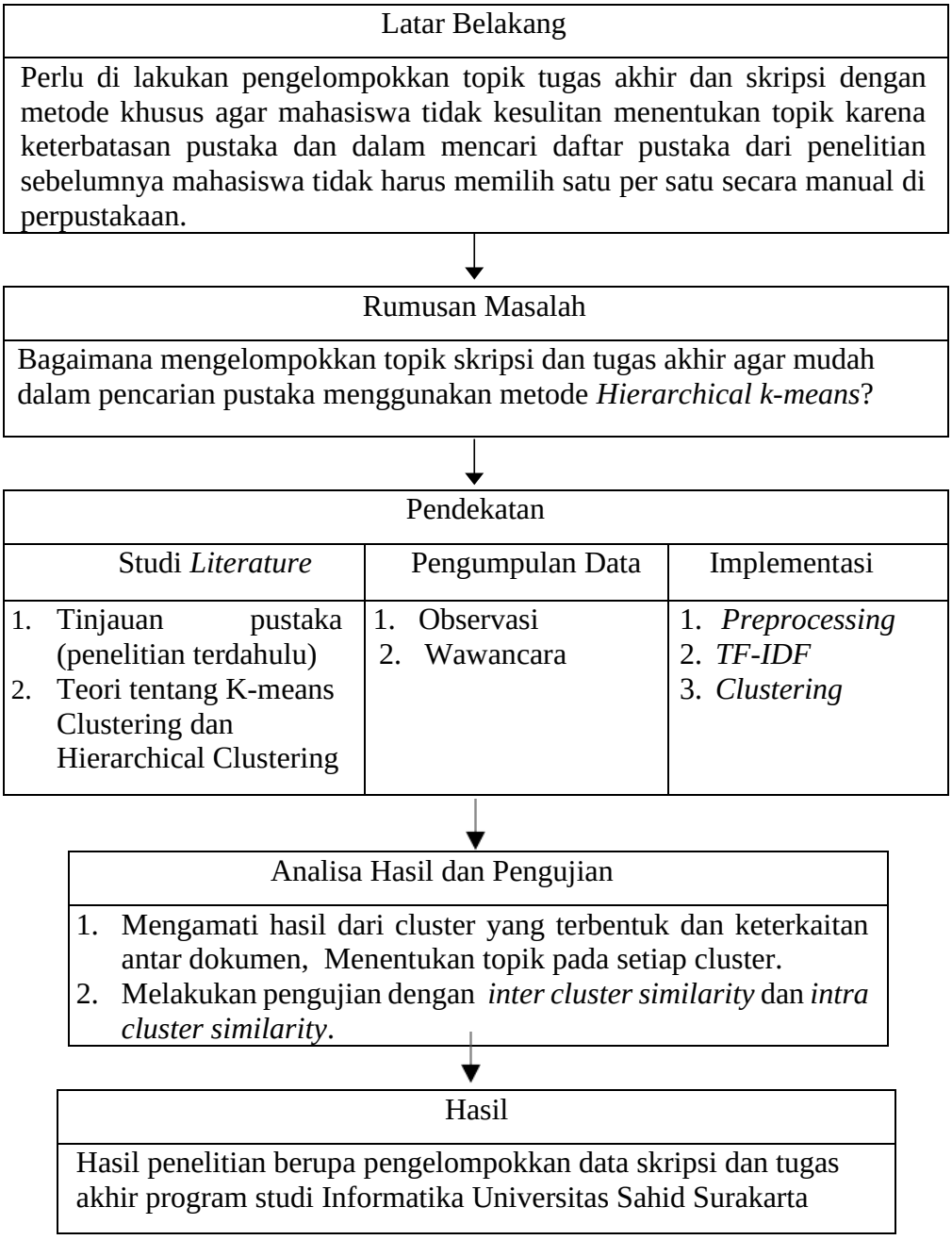
No	Nama, tahun, Judul	Rumusan masalah	Hasil
1	Tahta Alfina, Budi Santosa, dan Ali Ridho Barakbah (2012) "Analisa Perbandingan Metode Hierarchical Clustering, K-means dan Gabungan Keduanya dalam Cluster Data (Studi kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS)"	Bagaimana cara mengelompokkan problem kerja praktek berdasarkan posting problem yang ada pada forum diskusi online SI-KP yang ada di jejaring sosial facebook	Kombinasi antara metode Hierarchical clustering dengan metode K-means Clustering menghasilkan pengelompokan data yang lebih baik dibanding dengan yang hanya menggunakan metode K-means. Metode dari penelitian ini digunakan dalam proses clustering dokumen pada penelitian ini.
2	Lynda Rahmawat, Sari Widya Sihwi, Esti Suryani (2016) "Analisa clustering menggunakan metode k-means dan hierarchical clustering (studi kasus : dokumen skripsi jurusan kimia, fmipa, Universitas Sebelas Maret)"	Bagaimana melakukan pengelompokan terhadap dokumen skripsi Jurusan Kimia, Universitas Sebelas Maret dengan memanfaatkan proses data mining dengan menggunakan teknik clustering	Hasil analisa cluster menunjukkan bahwa penelitian di Jurusan Kimia pada tahun 2009-2012 terbatas pada beberapa tema. Tahun 2009 tema yang banyak diteliti adalah anorganik kompleks, organik dengan fokus antijamur dan antioksidan, dan kimia fisik. Tahun 2010 tema yang banyak diteliti adalah tema organik. Tahun 2011 tema yang diteliti adalah arganik dan anorganik. Pada tahun 2012 dan 2013 tema

No	Nama, tahun, Judul	Rumusan masalah	Hasil
			penelitian di Jurusan Kimia lebih bervariasi daripada tahun-tahun sebelumnya. Hasil analisa cluster dokumen skripsi Jurusan Kimia, FMIPA, UNS memperlihatkan bahwa keahlian dosen sangat mempengaruhi variasi tema penelitian yang dilakukan oleh mahasiswa. Variasi tema proyek dosen mempengaruhi variasi tema penelitian mahasiswa.
3	Rahmatika Diana Firdaus, Tri Ginanjar Laksana dan Rima Dias Ramadani (2019) "Pengelompokan Data Persedian Obat Menggunakan Perbandingan Metode K-Means dan Hierarchical Clustering Single Linkage (Studi Kasus di Puskesmas II Ajibarang)"	Bagaimana hasil klaster yang didapat dari keduanya dan algoritma mana yang terbaik dilihat dari hasil validitas klaster untuk studi kasus persediaan obat di Puskesmas II Ajibarang?	Algoritma K-Means dan HCC Single Linkage mampu mendapatkan klaster optimal sesuai dengan data kemiripan. Klaster optimal yang didapat dari kedua algoritma yaitu algoritma K-Means 180 data berada pada C1 atau obat dengan pemakaian lambat dan 24 data berada pada C2 atau obat dengan pemakaian cepat. Algoritma HCC Single Linkage dengan hasil 203 data berada pada C1 atau obat dengan pemakaian lambat dan 1 data berada pada C2 atau obat dengan pemakaian cepat. Nilai validitas algoritma k-Means dan HCC Single Linkage tergolong tinggi (strong structure) dibuktikan dengan pengujian validitas Sillhoutte Index masing-masing mendapatkan nilai validitas SI sebesar 0.8014 dan 0.8629. Oleh karena itu, nilai validitas terbaik berdasarkan pengujian SI diantara keduanya diperoleh algoritma HCC Single Linkage.

No	Nama, tahun, Judul	Rumusan masalah	Hasil
4	RA Pramudito (2021) "Clustering Berita Online Mengenai Covid-19 Menggunakan Hierarchical Clustering dan K-Means*"*	Bagaimana pengelompokan pemberitaan tentang Covid-19 setelah dilakukan text mining menggunakan gabungan metode hierarchical clustering dan k-means clustering?	Terbentuk cluster dengan 12 cluster dengan rata-rata intra cluster similarity 0.7911, rata-rata inter cluster similarity 0.3311.
5	Rendy Handoyo, R. Rumani M dan Surya Michrandi Nasution (2014) "Perbandingan metode clustering menggunakan metode single linkage dan k - means pada pengelompokan dokumen"	Bagaimana cara mengelompokkan dokumen berita berdasarkan tingkat kemiripan dari dokumen tersebut?	Nilai Silhouette Coefficient Single Linkage selalu lebih unggul dibandingkan dengan K-Means. Pertambahan jumlah dokumen membuat nilai Silhouette Coefficient single linkage semakin kecil sedangkan K-means terkadang menghasilkan nilai yang negatif. Untuk nilai Purity , Single Linkage selalu bernilai 1 sedangkan K-Means tidak pernah bernilai 1. Jadi metode Single Linkage memiliki performansi yang lebih baik dibandingkan dengan metode K-means
6	Riani (2021) "Prediksi Pendapatan Retribusi Uji KIR Menggunakan Algoritma C4.5 Pada Dinas Perhubungan Kabupaten Karanganyar"	Bagaimana prediksi pendapatan retribusi uji KIR menggunakan algoritma C4.5 pada dinas perhubungan kabupaten karanganyar?	Pendapatan retribusi Uji KIR untuk tahun 2021 adalah naik. Jenis kendaraan yang memiliki pengaruh besar dalam memprediksi kenaikan pendapatan retribusi Uji KIR adalah Light Truck, Pick Up, Mobil Baru dan Minibus. Nilai akurasi dari perhitungan secara manual dan RapidMiner menunjukkan nilai yang sama, sebesar 75%.

2.2 Kerangka Pemikiran

Kerangka pemikiran dari penelitian yang dilakukan dapat dilihat pada Gambar 2.1.



Gambar 2.1 Kerangka Pemikiran

Kerangka pemikiran dari penelitian yang dilakukan dapat diuraikan sebagai berikut:

1. Latar Belakang

   Pokok permasalahan yang mendasari penelitian ini adalah Perlu di lakukan pengelompokkan topik tugas akhir dan skripsi dengan metode khusus agar mahasiswa tidak kesulitan menentukan topik karena keterbatasan pustaka dan dalam mencari daftar pustaka dari penelitian sebelumnya mahasiswa tidak harus memilih satu per satu secara manual di perpustakaan.

2. Masalah

   Masalah yang akan diselesaikan dalam penelitian ini adalah Bagaimana mengelompokkan topik skripsi dan tugas akhir agar mudah dalam pencarian pustaka menggunakan metode *Hierarchical k-means*?

3. Pendekatan

   Pendekatan penelitian terdiri dari analisa studi *literature* tentang penelitian terdahulu dan kajian teori tentang K-means Clustering dan Hierarchical Clustering. Pengumpulan data TA dan skripsi mahasiswa prodi Informatika dan wawancara terhadap pustakawan kemudian ke tahap clustering yang terdiri dari 3 tahap yaitu *preprocessing*, *TF-IDF* dan *Clustering*.

4. Analisa Hasil dan Pengujian

   Mengamati hasil dari *cluster* yang terbentuk dan keterkaitan antar dokumen, menentukan topik pada setiap *cluster* kemudian melakukan pengujian menggunakan *inter cluster similarity* dan *intra cluster similarity*.

5. Hasil penelitian

   Berupa hasil pengelompokkan data skripsi dan tugas akhir program studi Informatika Universitas Sahid Surakarta.

2.3 Landasan Teori

2.3.1 *Data Mining*

Data Mining merupakan kompleks teknologi yang berakar pada berbagai disiplin ilmu matematika, statistik, ilmu komputer, fisika, teknik, biologi, dengan beragam aplikasi dalam berbagai macam domain yang berbeda seperti bisnis, kesehatan, sains dan teknik. Pada dasarnya, data mining dapat dilihat sebagai ilmu

menjelajahi dataset besar untuk mengekstraksi informasi tersirat, yang sebelumnya tidak diketahui dan berpotensi berguna (Harjanta, 2015).

Teknik data mining biasanya terbagi dalam dua kategori, prediksi dan deskripsi. Teknik prediksi menggunakan data historis untuk menyimpulkan sesuatu tentang kejadian di masa depan yang termasuk teknik ini adalah *clasification*, *regression*, *time series analysis* dan *prediction*. Sedangkan teknik deskripsi bertujuan untuk menemukan pola dalam data yang menyediakan beberapa informasi tentang hubungan interval yang tersembunyi yang termasuk teknik ini adalah *clustering*, *summarization*, *association rules*, *sequence discovery* (Nur Khormarudin, 2016)

2.3.2 *Clustering*

Clustering merupakan salah satu metode *Data Mining* yang bersifat tanpa arahan (*unsupervised*) (Rosmini dkk., 2018). *Clustering* adalah proses mengelompokkan atau penggolongan objek berdasarkan informasi yang diperoleh dari data yang menjelaskan hubungan antar objek dengan prinsip untuk memaksimalkan kesamaan antar anggota satu kelas dan meminimumkan kesamaan antar kelas *cluster* (Yudi Agusta, 2007).

Clustering merupakan salah satu teknik *data mining* yang digunakan untuk mendapatkan kelompok-kelompok dari objek-objek yang mempunyai karakteristik yang umum di data yang cukup besar. Tujuan utama dari metode *clustering* adalah pengelompokan sejumlah data atau objek ke dalam cluster atau grup sehingga dalam setiap cluster akan berisi data yang semirip mungkin (Nur Khormarudin, 2016).

2.3.3 *Text mining*

Text mining merupakan sebuah proses *unsupervised learning* untuk mengelompokkan kemiripan suatu dokumen dengan dokumen yang lain sehingga dapat dipisahkan menjadi beberapa kelompok (Handoyo dkk., 2014). Perbedaan mendasar dari *text mining* dan *data mining* terletak pada sumber data yang digunakan. Pada *data mining* data yang diekstrak berasal dari pola-pola tertentu dan terstruktur, sedangkam *text mining* sumber data yang digunakan berasal dari teks

yang relatif tidak testruktur karena menggunakan tata bahasa manusia atau biasa disebut *natural language* (Jumeilah, 2017).

Dalam penerapannya *text mining* dapat digunakan untuk *clasification, information extraction, information retrival* dan *clustering*(Firdaus & Firdaus, 2021).

2.3.3.1 *Clasification*

Klasifikasi adalah teknik yang paling umum diterapkan pada data mining. Pendekatan ini sering menggunakan keputusan pohon (*decision tree*) atau *neural network* berbasis algoritma klasifikasi. Proses klasifikasi data melibatkan learning dan klasifikasi. Dalam belajar (*learning*) data pelatihan (*training*) dianalisis dengan alogaritma klasifikasi. Dalam klasifikasi pengujian data dilakukan dengan menggunakan perkiraan akurasi dari aturan klasifikasi. Jika akurasi bisa diterima, maka aturan dapat diterapkan untuk data baru (Nur Khormarudin, 2016).

2.3.3.2 *Information extraction*

Ekstraksi informasi adalah cabang ruang lingkup dari text mining. Yang bertujuan untuk mengubah hasil proses text mining menjadi akar yang sama dengan dunia data yang terstruktur dalam data mining (Budiarti, 2006).

2.3.3.3 *Information retrival*

IR (*Information Retrieval*)merupakan suatu cara yang digunakan untuk menemukan kembali (*retrieve*) informasi-informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi secara otomatis (Pratama, 2018).

2.3.3.4 *Clustering*

Clustering merupakan proses membagi data dalam suatu himpunan ke dalam beberapa kelompok yang kesamaan datanya dalam suatu kelompok lebih besar daripada kesamaan data tersebut dengan data dalam kelompok lain (Maulida, 2018).

Clustering memiliki dua metode, yaitu *Hierarchical clustering dan partitioned clustering*. *Hierarchical clustering* mengelompokkan data secara bertahap, sedangkan *partitioned clustering* langsung mengelompokkan data

dengan menentukan jumlah *Cluster* di awal proses *clustering*. Salah satu metode *partitioned clustering* adalah *k-means clustering*. Penelitian ini menggunakan kombinasi antara *hierarchical clustering* dan *k-means clustering* (Alfina & Santosa, 2012).

2.3.4 *K-means Clustering*

K-Means merupakan metode penganalisaan data pada *data mining* dimana proses pemodelan tanpa supervisi dan merupakan salah satu metode yang mengelompokkan data secara partisi. Pada metode *K-Means* data dikelompokan menjadi beberapa kelompok dimana setiap kelompok mempunyai karakteristik yang mirip atau sama dengan lainnya namun dengan kelompok lainnya memiliki karakteristik yang berbeda. Metode ini menimalisasi perbedaan antar data di dalam satu *cluster* serta memaksimalkan perbedaan dengan *cluster* yang lain (Irfiani & Rani, 2018).

Metode *K-Means* memiliki karakteristik sebagai berikut:

1. *K-Means* merupakan metode pengelompokan yang sederhana dan dapat digunakan dengan mudah.
2. Pada jenis data set tertentu, *K-Means* tidak dapat melakukan segmentasi data dengan baik di mana hasil segmentasi tidak dapat menentukan pola kelompok yang mewakili karakteristik bentuk alami data.
3. K-Means biasa mengalami masalah ketika mengelompokkan data yang mengandung *outlier*.

Secara umum metode K-Means menggunakan algoritma sebagai berikut (Ediyanto dkk., 2013):

a. Tentukan k sebagai jumlah *cluster* yang di bentuk. Penentuan banyaknya jumlah cluster k dilakukan dengan beberapa faktor seperti pertimbangan teoritis dan konseptual yang diusulkan untuk menentukan berapa banyak *cluster*.
b. Bangkitkan k *Centroid* (titik pusat *cluster*) awal secara random. Untuk menentukan *centroid* awal dilakukan secara acak dari beberapa objek yang

tersedia sebanyak k *cluster*, untuk menghitung *centroid cluster* ke-i berikutnya, menggunakan rumus sebagai berikut:

$$v = \frac{\sum_{i=0}^{n} x_i}{n} \text{ ; i=1,2,3,..n}$$

(2. 1)

Keterangan simbol :

v : centroid pada cluster

xi : objek ke-i

n : banyaknya objek atau jumlah objek yang menjadi anggota *cluster*

c. Hitung jarak setiap objek ke masing-masing *centroid* dari masing-masing *cluster*. Kemudian hitung jarak antara objek dengan *centroid*, dalam penelitian ini menggunakan *Euclidian Distance*.

$$d(x_i, u_i) = \sqrt{(x_i - u_i)^2} \text{ ;}$$

(2. 2)

Keterangan simbol:

D : jarak antar *cluster*

xi : bobot kata ke i pada *cluster* yang ingin dicari jaraknya

ui: bobot kata ke i pada pusat *cluster*

d. Alokasikan masing-masing objek ke dalam *centroid* yang paling terdekat.

e. Lakukan iterasi, kemudian tentukan posisi *centroid* baru dengan menggunakan persamaan.

f. Ulangi langkah ke-3 jika posisi *centroid* baru tidak sama.

Proses penggabungan titik dilakukan dengan membandingkan matriks kumpulan tugas-tugas pada iterasi sebelumnya dengan matrik kumpulan tugas-tugas pada iterasi yang sedang berjalan. Jika hasilnya sama maka algoritma *kmeans cluster analysis* sudah konvergen, tetapi jika berbeda maka belum konvergen sehingga perlu dilakukan iterasi berikutnya.

2.3.5 *Hierarchical Clustering*

Hierarchical clustering adalah metode analisis kelompok yang berusaha untuk membangun sebuah hierarki kelompok. *Hierarchical clustering* dibagi

menjadi dua yaitu *Agglomeratif Clustering* dan *Difisive Clustering*. *Agglomeratif Clustering* mengelompokkan data dengan pendekatan bawah atas (*bottom up*), sedangkan *Difisive Clustering* menggunakan pendekatan atas bawah (*top-bottom*) (Rahmawati dkk., 2016).

Pada algoritma *hierarchical clustering* terdapat beberapa keunggulan yaitu tidak perlu menentukan jumlah *cluster* yang diinginkan karena proses dapat langsung dihentikan pada saat jumlah *cluster* sesuai dengan yang diinginkan. Namun algoritma ini juga memiliki kelemahan bergantung pada pemilihan teknik *intercluster similarity* yang lebih dikenal dengan istilah *linkage.* Beberapa kelemahan dari *linkage* tersebut adalah sensitif terhadap adanya *outlier,* kesulitan menangani variasi bentuk dan ukuran, dan memisahkan *cluster* yang besar. Tang, dalam jurnal (Zahrotun, 2015).

Kemiripan antar dokumen ditentukan dengan mengukur jarak antar dokumen. Dua dokumen yang mempunyai jarak paling kecil dikatakan mempunyai kemiripan paling tinggi dan dikelompokkan kedalam satu *cluster* yang sama. Sebaliknya dua dokumen yang mempunyai jarak paling besar dikatakan mempunyai kemiripan paling rendah dan dimasukkan ke dalam *cluster* yang berbeda. Langkah-langkah dalam algoritma *Hierarchical Agglomerative Clustering* (Arifin dkk., 2017):

1. Meletakkan setiap data sebagai sebuah *cluster*.
2. Hitung jarak matrik.
3. Gabungkan dua *cluster* yang paling dekat sesuai dengan parameter yang dipilih seperti *single, complete, average*.
4. Memperbarui jarak matrik dari *cluster* baru dengan *cluster* yang tersisa.
5. Lakukan kembali langkah 3 dan 4 sampai yang tersisa hanya satu *cluster.*

Metode dalam *Hierarchical Clustering* yang digunakan untuk menentukan *centroid* awal yaitu dengan *single linkage*.

2.3.6 Proses *text mining*

Pada proses *text mining* ada 2 tahapan yang harus dilakukan sebelum masuk ke proses *clustering*, yaitu *preprocessing* dan *TF-IDF*.

a. *Text preprocessing*

Text Preprocessing menjadi tahap awal dalam klasifikasi teks untuk mempersiapkan data teks sebelum digunakan pada proses lainnya. Pada tahap ini akan mengubah data teks menjadi bentuk yang lebih baik sehingga menghasilkan informasi teks dengan kualitas yang baik dan siap digunakan pada proses selanjutnya(Khairunnisa dkk., 2021). Terdapat beberapa tahapan untuk memproses data teks tersebut, yaitu :

1. *Case folding*

   Tahap awal adalah *case folding* yang bertujuan untuk mengubah setiap bentuk kata menjadi sama. Hal ini dilakukan dengan mengubah kata menjadi *lower case* atau huruf kecil. Karakter yang selain huruf tersebut akan dihilangkan dan menghilangkan karakter tidak valid seperti angka, tanda baca serta *Uniform Resources Locator* (Luqyana, 2018).

2. *Filtering*

   Filtering adalah sebagai proses mengambil kata – kata penting dari hasil proses *token* atau penghapusan *stopwords*. Bisa menggunakan algoritma *stop list* atau *word list. Stoplist* atau *stopword* merupakan kata-kata non-deskriptif dan dapat dibuang menggunakan pendekatan *bag-of-words* (Ratniasih dkk., 2017).

3. *Stemming*

   Stemming adalah suatu teknik pencarian bentuk dasar dari suatu *term*. Yang dimaksud dengan *term* itu sendiri adalah tiap kata yang berada pada suatu dokumen teks (Wibowo, 2016).

4. *Tokenization*

   Tokenization merupakan tahapan penguraian string teks menjadi *term* atau kata. Tujuan dari *Tokenization* yaitu memisahkan kata-kata dalam sebuah paragraf, kalimat atau halaman ke dalam kata tunggal (Najjichah dkk., 2019).

b. TF-IDF *(Term Frequency-Inverse Document Frequency)*

Term Frequency adalah salah satu metode yang digunakan untuk menghitung bobot tiap *term* dalam *text*. Nilai kepentingan setiap *term*, dalam

metode ini, dianggap sebanding dengan jumlah kemunculan term tersebut pada teks(Langgeni dkk., 2010). Berikut rumus untuk menghitung TF:

$$W_{(d,t)} = Tf_{(d,t)} \tag{2. 3}$$

$Tf_{(d,t)}$ merupakan *term frequency* dari *term* t dalam *text* d. Nilai *recall* pada *information retrieval* dapat diperbaiki oleh *term frequency*, namun nilai *precision* tidak selalu dapat diperbaiki. Hal ini dikarenakan kecenderungan *term* yang *frequent* muncul di banyak teks, sehingga kekuatan *diskriminatif* (keunikan) yang dimiliki *term-term* tersebut itu kecil. Permasalahan ini dapat diatasi dengan membuang *term* dengan nilai frekuensi yang tinggi dari *set term*. Fokus dari metode ini adalah untuk menemukan *threshold* yang optimal.

Sedangkan IDF (*Inverse Document Frequency*) merupakan frekuensi kemunculan *term* di pada keseluruhan dokumen. Nilai IDF berkaitan dengan distribusi *term* di berbagai dokumen. *Term* yang jarang muncul pada keseluruhan dokumen memiliki nilai IDF lebih besar dibandingkan dengan *term* yang bersangkutan, maka nilai IDF dari *term* tersebut adalah nol. Hal tersebut menunjukkan bahwa setiap *term* yang muncul pada dokumen dalam koleksi tidak berguna untuk membedakan dokumen berdasarkan topik tertentu (Manning, dkk., 2008).

Berikut rumus untuk menghitung IDF:

$$\log \frac{D}{df} \tag{2. 4}$$

Keterangan:

D : total dokumen

Df : banyak dokumen yang mengandung term yang dicari

Penelitian akhir-akhir ini telah menggabungkan Tf dan Idf untuk menghitung *term weighting*. Hasil yang ditunjukkan dari gabungan kedua metode tersebut adalah lebih baik. Berikut kombinasi kedua metode tersebut dirumuskan

$$W_{dt} = tf_{dt} \times IDF_t$$

(2. 5)

Keterangan :

d : dokumen ke-d

t : kata ke-t dari kata kunci

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

tf : banyaknya term yang dicari pada sebuah dokumen

IDF : Inverse Document Frequency

2.3.7 Evaluasi *Clustering*

Evaluasi merupakan tahap untuk mengetahui sejauh mana kualitas dari penelitian yang dilakukan, yang merupakan interpretasi hasil pemodelan yang digunakan (syifa & Fahmi, 2021). Terdapat 2 kriteria untuk menghitung kualitas hasil *clustering*:

2.3.7.1 *Intra Cluster Similarity*

Digunakan untuk menghitung rata-rata kedekatan (jarak) antara satu anggota dengan satu anggota lainnya.

$$I = \frac{1}{N^2} \sum_{didj \neq j} cosSim(di, dj)$$

(2. 6)

Sama seperti menghitung *norm* dari *centroid cluster*, berikut adalah perhitungan kedekatan antar tiap pasang antar tiap pasangan ketika menghitung *centroid* menggunakan *mean*.

$$I = \frac{1}{N^2} \sum_{didj \neq j} cosSim(di, dj) = ||c^2||$$

(2. 7)

Normalisasi dalam perhitungan ini dapat dilakukan sesuai dengan banyaknya anggota *cluster*.

$$I' = \frac{|| c^2||}{N}$$

(2. 8)

Hasil evaluasi dari *intra cluster similarity* menyimpulkan bahwa semakin tinggi kemiripannya, maka semakin bagus kualitas *cluster* tersebut.

2.3.7.2 *Inter Cluster Similarity*

Digunakan untuk menghitung perbedaan antara satu *cluster* dengan *cluster* lainnya (jarak minimum antar *cluster*). *Inter cluster similarity* dapat dihitung menggunakan *cosine similarity* antara satu *centroid cluster* dengan seluruh *centroid* suatu data.

$$E = \sum_{k=1}^{K} N_k \frac{C_k C^D}{||C_k||} \tag{2. 9}$$

Keterangan :

c^k : centroid *cluster* ke-k

c^D : centroid seluruh data

N_k : banyaknya *cluster*

Hasil evaluasi dari *inter cluster similarity* menyimpulkan bahwa semakin kecil nilainya, maka semakin bagus kualitas *clustering* tersebut.

2.3.8 *Flowchart*

Flowchart merupakan penyajian yang sistematis tentang proses dan logika dari kegiatan penanganan informasi atau penggambaran secara grafik dari langkah-langkah dan urut-urutan prosedur dari suatu program (Rejeki & Tarmuji, 2013).

Flowchart dibagi menjadi 5 jenis, yaitu (Regina, 2018):

1. *System flowchart*

   System Flowchart dapat didefinisikan sebagai bagan yang menunjukan arus pekerjaan secara keseluruhan dari sistem. Bagan ini menjelaskan urutan dari prosedur-prosedur yang ada di dalam sistem. Bagan alir sistem menunjukan apa yang dikerjakan dalam sistem.

2. *Document flowchart*

   Bagan alir dokumen (*document flowchart*) atau disebut juga bagan alir formulir (*form flowchart*) atau *paperwork flowchart* merupakan bagan alir yang menunjukan arus dari laporan dan formulir termasuk tembusan-tembusannya.

3. *Schematic flowchart*

   Bagan alir skematik (*schematic flowchart*) merupakan bagan alir yang mirip dengan bagan alir sistem, yaitu untuk menggambarkan prosedur di dalam sistem. Perbedaannya adalah bagan alir skematik selain menggunakan simbol-simbol bagan alir sistem, juga menggunakan gambar-gambar komputer dan peralatan lainnya yang digunakan. Penggunaan gambar-gambar ini digunakan untuk memudahkan komunikasi kepada orang yang kurang paham dengan simbol-simbol bagan alir. Penggunaan gambar-gambar ini mudah untuk dipahami tetapi sulit dan membutuhkan waktu lama untuk membuatnya.

4. Program *flowchart*

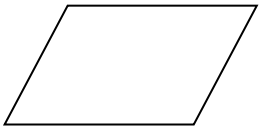
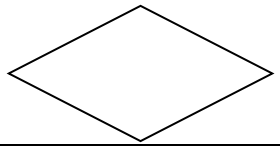
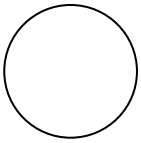
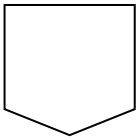
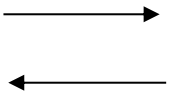
   Bagan alir program (*program flowchart*) merupakan bagan yang menjelaskan secara rinci langkah-langkah dari proses program. Bagan alir program dibuat dari derivikasi bagan alir sistem. Bagan alir program dapat terdiri dari dua macam, yaitu bagan alir logika program (*program logic flowchart*) dan bagan alir program komputer terinci (*detailed computer program flowchart*). Bagan alir logika program digunakan untuk menggambarkan tiap-tiap langkah di dalam program komputer secara logika. Bagan alat logika program ini dipersiapkan oleh analis sistem. Gambar berikut menunjukan bagan alir logika program. Bagan alir program komputer terinci (*detailed computer program flowchart*) digunakan untuk menggambarkan instruksi-instruksi program komputer secara terinci.

5. *Process flowchart*

   Bagan alir proses (*process flowchart*) merupakan bagan alir yang banyak digunakan di teknik industry. Bagan alir ini juga berguna bagi analis sistem untuk menggambarkan proses dalam suatu prosedur.

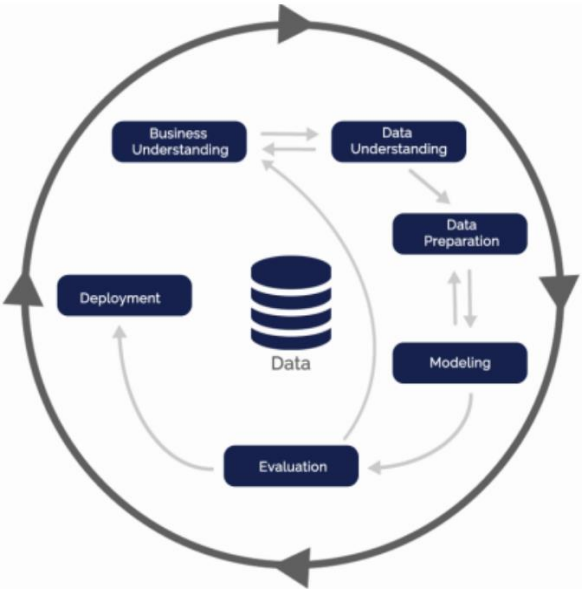
Dalam penggunaan *flowchart* terdapat simbol-simbol dasar yang sering digunakan yaitu (Khesya, 2021):

Tabel 2.2 Simbol *flowchart*

		
Proses	Input/Output	Keterangan
		
Pengujian	Pemberian nilai awal	Awal/akhir program
		
Konektor pada satu halaman	Konektor pada halaman lain	Arah

2.3.9 Metode CRISP-DM

Cross Industry Standard Process Model for Data Mining (CRISP-DM) merupakan proses strategi dalam pemecahan masalah secara umum dari bisnis atau unit penelitian(Nurchalifatun, 2017). Alur metode CRISP-DM dapat dilihat pada Gambar 2.2.



Gambar 2.2 Alur CRISP-DM (*Sumber*: Suhanda dkk., 2020)

CRISP-DM memiliki 6 tahapan(Feblian & Daihani, 2017):

1. *Business Understanding*

Business Understanding adalah pemahaman tentang substansi dari kegiatan *data mining* yang akan dilakukan, kebutuhan dari perspektif bisnis. Kegiatannya antara lain menentukan sasaran atau tujuan bisnis, memahami situasi bisnis, menerjemahkan tujuan bisnis kedalam tujuan data mining.

2. *Data Understanding*

Data Understanding adalah pengumpulan data, mempelajari data untuk dapat memahami data yang akan digunakan dalam penelitian, mengidentifikasi masalah yang berkaitan dengan data.

3. *Data Preparation*

Data Preparation, pada tahap ini struktur basis data akan dipersiapkan sehingga mempermudah proses *mining.*

4. *Modelling*

Modelling adalah tahap menentukan teknik data mining yang digunakan, menentukan *tools data mining*, algoritma *data mining*, menentukan parameter dengan nilai yang optimal.

5. *Evaluation*

Evaluation adalah tahap interpretasi terhadap hasil data mining yang ditunjukkan dalam proses pemodelan yang terdapat pada tahap sebelumnya. Evaluasi dilakukan secara mendalam dengan tujuan menyesuaikan model yang didapat agar sesuai dengan sasaran yang ingin dicapai dalam tahap pertama.

6. *Deployment*

Tahap *deployment* atau tahap penyebaran adalah tahap penyusunan laporan atau presentasi dari pengetahuan yang didapat dari evaluasi pada proses *data mining.*