BAB 2 LANDASAN KEPUSTAKAAN

2.1 Kajian Kepustakaan

Penelitian ini didasarkan pada studi literatur buku dan beberapa penelitian sebelumnya yang berkaitan dengan penerapan *information gain* untuk *reduksi* fitur dan *k-means* untuk pengelompokan dokumen. Publikasi penelitian terdahulu dalam bentuk paper, jurnal, skripsi digunakan sebagai sumber kajian pustaka dalam penelitian ini. Studi literatur ini dilakukan untuk memperkuat pemahaman mengenai permasalahan dan menemukan solusi terbaik.

Pengelompokan dokumen secara luas diklasifikasikan menjadi dua kelompok yaitu *hard clustering* dan *soft clustering. Hard Clustering* adalah pengelompokan dokumen dimana dokumen dimana setiap dokumen hanya memiliki satu kelompok. Sedangkan *Soft Clustering* adalah pengelompokan dokumen dimana satu dokumen dapat memiliki lebih dari satu kelompok (Zade, dkk, 2017).

Salah satu penelitian terdahulu yang membahas mengenai pengelompokan dokumen adalah penelitian milik Zaini, dkk (2017). Pada penelitian Zaini, dkk (2017) digunakan data berupa artikel berbahasa Indonesia dan metode Self Organizing Map, dimana urutan kemunculan kata sangat diperhatikan. Sebelum dikelompokan, dilakukan *pre-processing* terhadap artikel. Tujuan dari pre-processing adalah untuk membuang kata yang tidak penting dan mempermudah ekstraksi fitur. Selanjutnya dilakukan pembobotan *TF-IDF* yaitu proses menentukan bobot kata berdasarkan *frekuensi* kemunculan.

Selanjutnya ada penelitian dari Wibowo, dkk(2017) dengan menggunakan data berupa tugas akhir dan skripsi. Jika dalam penelitian Zaini, dkk setelah tahap pembobotan *TF-IDF* dilanjutkan dengan mencari struktur laten kemunculan kata baru dikelompokan, pada penelitian Wibowo, dkk, 2017 langsung dilakukan pengelompokan dokumen dengan metode *Cosine Similarity*. Metode *Cosine Similarity* digunakan untuk menentukan kedekatan jarak dokumen satu dengan yang lainnya. Hasilnya dari penelitian menunjukan bahwa dokumen dapat dikelompokan berdasarkan kategori yang dijadikan kata kunci (Wibowo, dkk, 2017).

Selanjutnya ada penelitian milik Zade, dkk, 2017. Data yang digunakan dalam penelitian tersebut berupa dokumen teks. Didalam penelitian tersebut pengolahan teks tidak dijelaskan, dianggap semua dokumen telah memiliki jarak satu sama lain. Metode yang digunakan adalah metode *K-Means.* Pertama ditentukan dahulu k kelompok kemudian menentukan nilai tengah dari setiap kelompok. Setiap dokumen yang dekat dengan nilai tengah dari kelompok tertentu akan masuk ke dalam kelompok tersebut. Hasil dari penelitian ini menunjukan semakin banyak dokumen maka semakin akurat hasilnya. (Zade, dkk, 2017)

Dari ketiga penelitian tersebut, setiap metode memiliki kelebihan dan kekurangan masing masing. Untuk mendapatkan hasil yang optimal maka dari ketiga metode diatas dapat diambil satu teknik yang kemudian dapat digabungkan untuk membuat sistem pengelompokan yang akurat. Salah satu penelitian yang memanfaatkan teknik pre-

processing dan penentuan jarak dengan *cosine similarity* sebagai dasar pengelompokan dengan metode *K-Means* adalah penelitian dari Subandi.

Dalam penelitiannya data yang digunakan Subandi adalah dokumen skripsi. Sebelum dikelompokan dokumen skripsi di *pre-processing* kemudian hasil dari *pre-processing* ditentukan bobot kata nya. Dari bobot kata kemudian dihitung cosine similaritynya untuk menentukan kedekatan dokumen. Jarak kedekatan tersebut yang kemudian digunakan untuk jarak dalam pengelompokan. (Subandi, 2014)

Penelitian yang serupa juga dilakukan oleh Dewi. Dalam penelitiannya, Dewi menambahkan reduksi fitur *Information Gain Thresholding* guna mengurangi fitur yang ada sehingga komputasi akan lebih cepat dan menghilangkan noise. (Dewi, 2013). Berdasarkan beberapa referensi tersebut, maka pada penelitian kali ini akan digunakan beberapa ilmu yang berkaitan dengan *pre-processing, cosine similiraty, k-means* guna mendukung penelitian ini.

2.2 Pengelompokan *(Klastering)*

Pengelompokan adalah proses mengelompokan objek yang memiliki kesamaan ke dalam suatu kelompok yang memainkan peran yang penting bagi manusia untuk menganalisis dan menggambarkan kumpulan objek tersebut (Tan,dkk, 2006). Menurut Tan, dkk (2006) kegunaan dari pengelompokan adalah untuk peringkasan, kompresi, dan menemukan objek terdekat. Pengelompokan dibagi menjadi beberapa tipe, seperti :

- Hirarki dan Partisi

  Hirarki adalah pengelompokan dimana terdapat kelompok bersarang yang digambarkan seperti pohon. Setiap kelompok dalam pohon merupakan gabungan dari kelompok *root* dan kelompok *children*. Sehingga setiap objek pasti merupakan anggota dari kelompok *root*. Sedangkan partisi adalah pengelompokan dimana objek akan dikelompokan ke dalam kelompok *non-overlapping*, artinya setiap objek hanya akan memiliki satu kelompok yang berdiri sendiri.

  Dalam hal ini dapat disimpulkan bahwa pengelompokan hirarki merupakan rangkaian kelompok dari pengelompokan partisi. Sedangkan pengelompokan partisi merupakan kelompok dasar pada pengelompokan hirarki.

- Eksklusif, *overlapping*, dan *fuzzy*

  Pengelompokan eksklusif merupakan pengelompokan dimana objeknya dapat memiliki kelompok lebih dari satu. Berbeda dengan pengelompokan eksklusif, pengelompokan non-eksklusif atau *overlapping* adalah pengelompokan dimana objeknya hanya memiliki satu kelompok. Dalam beberapa penelitian, seperti penelitian Bora dan Gupta (2014) pengelompokan eksklusif disebut juga sebagai *hard clustering* dan pengelompokan overlapping disebut juga sebagai *soft clustering*. (Bora dan Gupta, 2014).

  Sedangkan pada pengelompokan *fuzzy* keanggotaan objek ditentukan berdasarkan nilai 0 hingga 1, dimana setiap kelompok dibentuk dari persamaan *fuzzy.*

- Lengkap dan Sebagian

  Pengelompokan lengkap merupakan pengelompokan dimana setiap objek yang akan dikelompokan memiliki kelompok. Sedangkan pengelompokan sebagaian merupakan pengelompokan dimana objeknya bisa tidak memiliki kelompok.

Pengelompokan memiliki hubungan yang erat dengan masalah pengurangan dimensi. Data berdimensi tinggi sering kali menantang untuk dianalisis, karena semakin meningkatnya keberagamanan data. Metode klastering dapat dilihat sebagai perpaduan antara metode seleksi fitur / *dimensionality reduction* dengan pengelompokan (Aggarwal dan Reddy, 2014).

2.3 *Text Mining*

Text mining memiliki definisi menggali data berupa teks yang sumbernya berupa dokumen untuk mencari kata kata yang dapat mewakili dokumen. Ilmu *Text Mining* digunakan untuk mengubah kumpulan teks menjadi numerik sehingga dapat dikomputasikan. Teknik ini disebut sebagai teknik *pre-processing* (Sari dan Puspaningrum, 2013)

2.3.1 Pre- Processing Teks

Pre-processing text merupakan tahapan awal dalam text mining dimana tujuannya adalah melakukan pembersihan terhadap kata kata yang tidak penting sehingga kata yang berkualitas dapat diproses menggunakan algoritma tertentu (Sanjaya dan Absar, 2015). Tahapan pre-processing meliputi *tokenisasi, stopword removal,* dan *stemming* (Zaini, dkk, 2017).

2.3.1.1 Tokenisasi

Tokenisasi adalah proses memotong kalimat menjadi potongan- potongan kata, yang disebut token, dan pada saat yang sama karakter-karakter tertentu, seperti tanda baca dihapus (Manning, 2008). Singkatnya *tokenisasi* adalah proses memisahkan deretan kata menjadi potongan kata yang memiliki makna.

2.3.1.2 Stopword Removal

Stopword adalah pembuangan kata yang sering muncul tapi tidak memiliki makna yang penting (Kogilavani dan Balasubramani, 2010). Singkatnya pada tahapan *stopword removal*, kata kata tidak penting yang dihilangkan. Kata kata tidak penting disini maksudnya adalah kata kata yang kurang memiliki makna dan terlalu sering muncul seperti kata kata konjungsi seperti yang, di, dan, dan sebagainya.

2.3.2 Stemming

Stemming merupakan teknik untuk mengubah token menjadi kata dasar. Kata dasar biasanya digunakan di beberapa artikel dengan berbahagai imbuhan yang bisa jadi sama atau berbeda (Zaini, dkk, 2017). Proses *stemming* dalam bahasa Indonesia lebih kompleks, karena terdapat berbagai macam variasi serta kombinasi imbuhan yang harus dihapus untuk mendapatkan kata dasar (Abdurrasyid, 2012). Dalam Penelitian ini digunakan *stemmer* dari lucene-solr atau IndonesianStemmer.java. Lucene-solr merupalan library

yang disediakan oleh java. Dalam penerapannya, lucene-solr harus disimpan dalam library kemudian untuk menggunakannya perlu dilakukan *import library.*

2.3.3 Pembobotan TF-IDF

Term Frequency dan Inverse Document Frequency (TF-IDF) merupakan pembobotan yang sering digunakan dalam penelusuran informasi dan text mining (Turney dkk, 2010). Term frequency adalah pembobotan yang sederhana dimana penting tidaknya sebuah kata dianggap sama atau sebanding dengan jumlah kemunculan kata tersebut dalam dokumen, sementara itu inverse document frequency (IDF) adalah pembobotan yang mengukur penting sebuah kata dalam dokumen dilihat pada seluruh dokumen secara global. Disini dokumen dan query yang telah ditokenisasi, filtering, dan stemming dihitung bobotnya.

Mencari nila term-frequency melalui persamaan 2.1

$$Tf_{t,d} = 1 + {}^{10}\text{Log } tf \qquad (2.1)$$

Dimana setiap variable dijelaskan sebagai berikut :

tf : term frekuensi atau banyaknya kata pada dokumen

$Tf_{t,d}$: term frekuensi atau banyaknya kata t pada dokumen d atau pembobotan local

Mencari nilai inverse document- frequency melalui persamaan 2.2

$$idf_t = {}^{10}\log {}^{n}/_{df_t} \qquad (2.2)$$

Idf_t : inverse document- frequency atau pembobotan global

n : banyaknya dokumen

df_t : banyaknya dokumen yang memiliki kata t

Dari persamaan 2.1 dan 2.2 baru bisa ditentukan nilai bobotnya ($W_{t,d}$) dengan mengalikan kedua persamaan sehingga menjadi persamaan 2.3

$$W_{t,d} = tf_{t,d} \times idf_t \qquad (2.3)$$

$Tf_{t,d}$: term frekuensi atau banyaknya kata pada dokumen atau pembobotan local

Idf_t : *inverse document frequency* atau pembobotan global

$W_{t,d}$: nilai bobot akhir kata

Kemudian lakukan normalisasi pada bobot yang telah didapatkan dengan menggunakan persamaan 2.4 :

$$\frac{W_{t,d}}{\sqrt{\sum_{t=1}^{n}(W_{t,d})^2}} \quad (2.4)$$

Dimana :

n : banyaknya kata
t : iterasi kata ke-
Wt,d : nilai bobot akhir kata

Pembobotan tf-idf digunakan untuk mengukur seberapa penting suatu kata dalam suatu dokumen. Untuk perhitungan tf-idf pada dapat dilihat pada implementasi.

2.3.4 Cosine Similarity

Dalam tahap ini kemiripan dokumen skripsi dengan setiap dokumen yang ada dihitung. Hitung kemiripan vektor query Q dengan setiap dokumen yang ada. Kemiripan antar dokumen dapat menggunakan cosine similarity. Rumus di tuliskan pada persamaan 2.5.

$$\cos\theta_{ki} = \sum_{k}(d_{ik}d_{jk}) \quad (2.5)$$

Dimana diketahui nilai variabel dari persamaan 2.5

k : jumlah dokumen
d_{ik} : panjang dokumen ke i
d_{jk} : panjang dokumen ke query
$\cos\theta_{kj}$: kedekatan dokumen yang dicari

2.4 *Data Mining*

2.4.1 Reduksi Fitur dengan *Information Gain Thresholding*

Information Gain merupakan teknik *reduksi fitur* menggunakan metode pembobotan atribut kontinu yang dideskretkan dengan maksimal entropy atau nilai *information gain*. *Entropy* digunakan untuk mendefinisikan nilai *Information Gain*. *Entropy* menggambarkan banyaknya informasi yang dibutuhkan untuk mengkodekan suatu kelas. *Information Gain (IG)* dari suatu term diukur dengan menghitung jumlah bit informasi yang diambil dari prediksi kategori dengan ada atau tidaknya term dalam suatu dokumen (Maulida, dkk, 2016). *Information Gain* atau biasa disebut IG adalah salah satu atribut pengukuran seleksi data untuk memilih tes pada atribut. Secara matematis dituliskan pada persamaan 2.6

$$\text{Entropy(S)} = -\Sigma\frac{|S_i|}{s}\,log\frac{|S_i|}{s} \quad (2.6)$$

Kemudian dari rumus *entropy* diatas dapat dicari nilai *Information Gain* dengan persamaan 2.7

InfoGain (S,A) = Entropy(S) - Σ|Sv|$Sv \in Value(A)\ Entropy(Sv)$ (2.7)

Dimana S adalah jumlah seluruh fitur, A adalah kategori, Sv adalah jumlah sampel untuk nilai v, v adalah nilai yang mungkin untuk kategori A, Si adalah fitur ke I, dan Value(A) adalah himpunan nilai-nilai yang mungkin untuk kategori A

Fitur yang dipilih adalah fitur dengan nilai Information Gain yang tidak sama dengan nol dan lebih besar dari suatu nilai threshold tertentu. Ide dibalik Information Gain untuk memilih fitur adalah menyatakan fitur dengan informasi yang paling signifikan terhadap kategori.

2.4.2 K-Means

Metode *k-means* adalah algoritma pembelajaran tanpa data latih. Metode ini paling sederhana dan paling banyak digunakan. Prosedur pengelompokan dari metode ini adalah mendefiniskan kelompok k dan satu k center untuk setiap cluster. (Zade, dkk, 2017)

Algoritma K-Means merupakan algoritma untuk mengelompokan dokumen berdasarkan jarak terdekat. Menurut Adiningsih (2007), tahap penyelesaian algoritma K-Means adalah sebagai
berikut:

1. Menentukan K buah titik yang merepresentasikan obyek pada setiap cluster (centroid awal).
2. Menetapkan setiap objek pada cluster dengan posisi centroid terdekat. Adapun cara untuk menentukan jarak yaitu dengan menggunakan persamaan 2.5.
3. Jika semua objek sudah dikelompokkan maka dilakukan perhitungan ulang dalam menentukan centroid yang baru. Untuk menentukan centroid baru persamaan yang digunakan dituliskan pada persamaan 2.9.

   $$Ci = \frac{x_i + \cdots + x_n}{\sum x} \quad (2.9)$$

   Keterangan :

   x1 = nilai data record ke-1

   x2 = nilai data record ke-2

   Σx = jumlah data record

4. Ulangi langkah 2 dan 3 sampai centroid tidak berubah