

## PERBANDINGAN PERINGKASAN MULTI DOKUMEN ILMIAH BERBAHASA INDONESIA MENGGUNAKAN METODE *K-MEANS* DAN *K-NEAREST NEIGHBORS (K-NN)*

Roza Yuniza Putri<sup>1</sup>, Widodo<sup>2</sup>, Hamidillah Ajie<sup>3</sup>

<sup>1</sup> Mahasiswa Prodi Pendidikan Teknik Informatika dan Komputer, Teknik Elektro, FT – UNJ

<sup>2,3</sup> Dosen Prodi Pendidikan Teknik Informatika dan Komputer, Teknik Elektro, FT – UNJ

[rozayuniza20@gmail.com](mailto:rozayuniza20@gmail.com), [widodo@unj.ac.id](mailto:widodo@unj.ac.id), [hamidillah@unj.ac.id](mailto:hamidillah@unj.ac.id)

---

### Abstrak

Dalam menyusun penelitian, dokumen ilmiah dibutuhkan sebagai salah satu sumber data yang dapat dipertanggungjawabkan. Namun dalam menentukan informasi utama, pembaca mengalami kesulitan untuk mendapatkan kesimpulan dari beberapa dokumen ilmiah. Peringkasan multi dokumen ilmiah berfungsi untuk memudahkan pembaca dalam memahami dan mendapatkan kesimpulan dari beberapa dokumen ilmiah. Ada banyak metode yang bisa digunakan untuk peringkasan multi dokumen ilmiah, salah satunya adalah metode *K-Means* dan *K-Nearest Neighbors (K-NN)*. Penelitian ini bertujuan untuk mengetahui dan menganalisa perbandingan kinerja peringkasan multi dokumen ilmiah berbahasa Indonesia menggunakan metode *K-Means* dan *K-Nearest Neighbor (K-NN)*. Bahan penelitian yang digunakan adalah dokumen ilmiah berbahasa Indonesia sebanyak 30 dokumen dengan topik yang sama tentang “Sistem Informasi Perpustakaan”. Dokumen ilmiah diproses menggunakan tools NLTK dan Sastrawi, kemudian hasil peringkasan metode *K-Means* dan *K-Nearest Neighbors (K-NN)* dievaluasi dengan metode ROUGE. Pengujian metode *K-Means* pada ROUGE-1 dan ROUGE-2 menunjukkan hasil optimal sebanyak 30 dokumen dengan tingkat peringkasan sebanyak 15 kalimat. Pada ROUGE-1 diperoleh nilai presisi 0.391, recall 0.277, dan *f-measure* 0.324. Kemudian pada ROUGE-2, diperoleh nilai presisi 0.067, recall 0.048, dan *f-measure* 0.056. Sedangkan pada metode *K-Nearest Neighbors (K-NN)* pada ROUGE-1 dan ROUGE-2 menunjukkan hasil optimal sebanyak 10 dokumen dengan tingkat peringkasan 15 kalimat. Pada ROUGE-1 diperoleh nilai presisi 0.362, recall 0.296, dan *f-measure* 0.326. Kemudian pada ROUGE-2, diperoleh nilai presisi 0.045, recall 0.037, dan *f-measure* 0.045. Dapat disimpulkan bahwa metode *K-Means* memiliki kinerja yang lebih baik jika dokumen yang digunakan relatif lebih banyak. Sebaliknya, metode *K-Nearest Neighbors (K-NN)* memiliki kinerja yang lebih baik jika dokumen yang digunakan relatif lebih sedikit.

**Kata kunci :** Peringkasan Multi Dokumen, Dokumen Ilmiah, *K-Means*, *K-Nearest Neighbors*, ROUGE

---

### 1. Pendahuluan

Salah satu pemanfaatan internet sebagai sumber informasi adalah dengan pencarian dokumen ilmiah untuk penyusunan karya ilmiah. Internet dapat menjadi sumber informasi dikarenakan informasi yang terdapat pada internet relatif lebih *up-to-date*. Berdasarkan data RISTEKDIKTI (2019) aksesibilitas halaman jurnal mengalami peningkatan dari waktu ke waktu, *visibility* dan dampak ilmiah di mesin pencari Google Scholar meningkat. Hal ini membuktikan bahwa dari waktu ke waktu banyak pembaca yang menjadikan jurnal ilmiah sebagai sumber utama maupun sumber penunjang dalam penyusunan karya ilmiah.

Dalam menyusun sebuah penelitian dibutuhkan data yang tepat dan akurat agar menjadi relevan. Dokumen ilmiah dibutuhkan dalam penyusunan sebuah penelitian karena merupakan salah satu sumber data yang dapat dipertanggungjawabkan. Namun, dalam menentukan informasi utama dari dokumen ilmiah, pembaca mengalami kesulitan karena pembaca harus memperdalam dan memahami dokumen tersebut untuk mendapatkan kesimpulan atau poin utama dari dokumen tersebut. Berdasarkan penelitian Aysiyah & Inawati (2020) bahwa pembaca memerlukan waktu yang relatif lama untuk menyaring informasi yang sesuai dengan kebutuhannya, hal ini dikarenakan informasi yang didapatkan dari Internet sangat banyak, sehingga pencarian informasi menjadi tidak maksimal. Oleh karena itu, diperlukan penelitian peringkasan teks untuk memperbaiki kualitas ringkasan agar lebih akurat dan memiliki koherensi kalimat yang makin kuat.

Berdasarkan uraian tersebut maka peringkasan multi dokumen ilmiah dalam kasus ini sangat dibutuhkan, hal ini bertujuan untuk menghasilkan sebuah ringkasan dari beberapa dokumen ilmiah agar membantu pembaca memperoleh informasi yang cepat sekaligus relevan tanpa harus membaca keseluruhan dari dokumen tersebut.

Ada beberapa metode yang dapat digunakan dalam peringkasan teks, salah satunya adalah metode *K-Means* dan *K-Nearest Neighbors* (K-NN). Berdasarkan hasil penelitian yang dilakukan oleh Husni (2015) dengan menggunakan metode *K-Means*, dokumen berita berhasil dikelompokkan secara otomatis sesuai dengan derajat kesamaan berita sehingga menjadi kelompok dokumen berita yang terstruktur. Selanjutnya, berdasarkan hasil penelitian yang dilakukan oleh Indrianto, Fauzi, & Muflikhah (2017), metode *K-Nearest Neighbors* (K-NN) berhasil diterapkan dalam peringkasan teks otomatis pada artikel berita kesehatan dengan melakukan *preprocessing*, ekstraksi fitur dan menghitung jarak *Euclidean Distance* untuk kelas terdekat. Berdasarkan hasil kedua penelitian yang telah dijabarkan sebelumnya, maka metode *K-Means* dan *K-Nearest Neighbors* (K-NN) memiliki kinerja yang cukup baik dalam melakukan peringkasan teks yang dibuktikan dengan berhasilnya peringkasan teks menggunakan metode tersebut. Namun, dalam peringkasan multi dokumen ilmiah perlu diketahui metode mana yang memiliki kinerja yang lebih baik.

## 2. Dasar Teori

Ada beberapa teori yang akan dijelaskan mengenai *dokumen ilmiah*, peringkasan teks, *text preprocessing*, pembobotan kalimat (TF-ISF), *cosine similarity*, metode *K-Means*, metode *K-Nearest Neighbors* (K-NN), dan *ROUGE*.

### 2.1. Dokumen Ilmiah

Dokumen ilmiah merupakan salah satu sumber data yang dapat dipertanggungjawabkan. Dokumen ilmiah itu sendiri ialah sebuah dokumentasi dari penelitian ilmiah yang dapat dijadikan referensi untuk penelitian ilmiah yang akan dilakukan. Dokumen ilmiah yang banyak dijadikan referensi diantaranya adalah skripsi, tesis, disertasi, jurnal, makalah ilmiah, dan lain – lain. Bahasa yang digunakan dalam dokumen ilmiah harus baik dan baku. Selain itu, bahasa yang digunakan harus diusahakan singkat, padat dan jelas. Tulisan ilmiah harus diusahakan memuat banyak informasi, sedikit kata-kata dan tidak bertele-tele (Anam, 2019)

### 2.2. Peringkasan Teks

Peringkasan teks merupakan proses mengambil intisari dari sebuah informasi sehingga teks tersebut menjadi lebih pendek dan mengacu pada poin – poin penting dari teks tersebut (Indriani 2014 diacu dalam Saraswati, Indriati, & Perdana, 2018). Dalam peringkasan teks otomatis, sistem diberikan sebuah teks, kemudian melakukan proses peringkasan, dan mengeluarkan hasil berupa ringkasan dari teks aslinya. Berdasarkan teknik peringkasan, ada dua teknik yang dapat digunakan yaitu ekstraksi dan abstraksi (Mustaqhfi, Abidin, & Kusumawati, 2011). Ekstraksi adalah penggalian beberapa kalimat penting dari dokumen teks asli yang memuat informasi penting dalam dokumen tersebut. Sedangkan abstraksi adalah pembuatan dan penggabungan kalimat baru, dimana kalimat baru tersebut merupakan inti dari informasi dalam dokumen ringkasan (Tardan, Erwin, Eng I, & Muliady, 2013). Dilihat dari jumlah sumber teks, proses ringkasan dapat diambil dari satu dokumen (*single document*) atau dari beberapa dokumen (*multi document*). Dalam ringkasan *single document*, masukannya adalah teks, dan keluarannya adalah teks yang lebih singkat dari teks sumber aslinya. Sedangkan pada *multi document*, masukannya adalah beberapa dokumen teks dengan jenis topik yang sama kemudian keluarannya adalah teks yang lebih singkat dan mencakup informasi utama (Pradnyana & Mogi, 2014). Peringkasan multi dokumen (*multi document*) berbeda dari peringkasan dokumen tunggal (*single document*). Dalam peringkasan *multi document*, ruang pencariannya lebih besar dibandingkan dengan peringkasan *single document* yang lebih sulit untuk mengekstrak kalimat penting. Dalam hal ini peringkasan multi dokumen dapat dianggap sebagai masalah optimasi, tujuannya untuk menghasilkan ringkasan optimal yang memuat kalimat informasi dari dokumen asli (Rautary & Balabantary, 2018).

### 2.3. Text Preprocessing

Salah satu tahapan penting dalam peringkasan teks adalah *text preprocessing*, dimana pada tahap ini teks asli diubah menjadi bentuk yang lebih pendek (Yuliska & Syaliman, 2020). Tujuan dari *text preprocessing* adalah mempersiapkan teks menjadi sebuah data, kemudian akan dilanjutkan ke proses berikutnya. Pada *text mining*, informasi yang dikumpulkan tidak berurutan. Oleh sebab itu, dibutuhkan proses penyusunan kembali agar menjadi data yang berurutan sesuai kebutuhan untuk proses dalam *data mining*. Setelah data tersusun maka data tersebut dapat menjadi sumber data yang akan diolah lebih lanjut (Feldman & Sanger, 2007 diacu dalam Jumeilah, 2017). Tahapan yang dilakukan oleh *text preprocessing* adalah pemecahan kalimat, *case folding*, *tokenizing*, *stopword removal/filtering*, dan *stemming*.

### 2.4. Pembobotan Kalimat (TF-ISF)

*Term Frequency – Inverse Sentence Frequency* (TF-ISF) merupakan salah satu skema dari pembobotan *Term Frequency – Inverse Document Frequency* (TF-IDF). Dimana *term frequency* (TF) adalah frekuensi kemunculan *term t* pada kalimat *s* dalam sebuah dokumen. Sedangkan *Inverse Sentence Frequency* (ISF) adalah perhitungan

bagaimana *term t* tersebar secara menyeluruh pada sekumpulan kalimat dalam sebuah dokumen. Pembobotan berdasarkan *Term Frequency – Inverse Sentence Frequency* (TF-ISF) dapat dirumuskan sebagai berikut (Abdullah & Faticah, 2018):

$$ISF_t = \log \frac{N}{SF_t} \quad (1)$$

$$W_{st} = tf_{st} \times ISF_t \quad (2)$$

Keterangan :

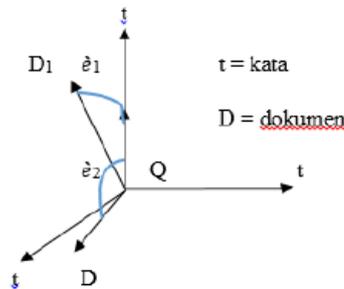
$tf_{st}$  = banyaknya kemunculan *term* ke-*t* pada kalimat *s*

$N$  = banyaknya kalimat dalam satu dokumen

$SF_t$  = banyak kalimat yang mengandung *term* ke-*t*

## 2.5. Cosine Similarity

*Cosine similarity* merupakan metode yang digunakan untuk menghitung nilai kemiripan antara dua *term*, yaitu dokumen dengan *query* atau dokumen dengan dokumen (Saraswati, Indriati, & Perdana, 2018). *Term* yang digunakan bisa berupa kata, paragraf, atau teks sumber secara keseluruhan. Setelah perhitungan selesai, bobot *cosine* akan diurutkan. Semakin banyak jumlah nilai *cosine* yang tinggi menentukan tingkat kemiripan yang tinggi (Purnamasari & Widiastuti, 2017).



Gambar 2. 1 Vektor Skalar *Cosine Similarity*

Persamaan *Cosine Similarity* dapat dirumuskan sebagai berikut:

$$\text{Cos}(b_1, b_2) = \frac{\sum_{t=1}^n W_{t.b1} W_{t.b2}}{\sqrt{\sum_{t=1}^n W^2_{t.b1}} \sqrt{\sum_{t=1}^n W^2_{t.b2}}} \quad (3)$$

Keterangan :

$\text{Cos}(b_1, b_2)$  = tingkat kesamaan dokumen dengan *query* tertentu

$W_{t.b1}$  = bobot *term* ke-*t* dalam vektor untuk dokumen  $b_1$

$W_{t.b2}$  = bobot *term* ke-*t* dalam vektor untuk dokumen  $b_2$

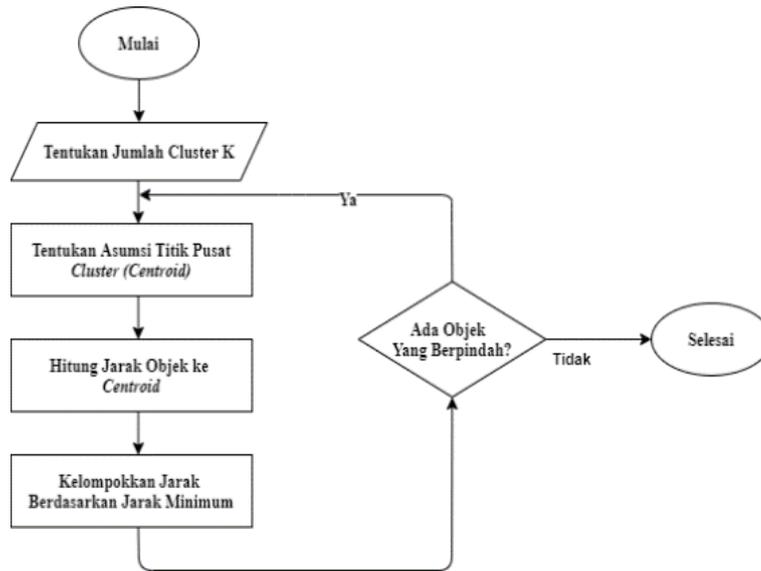
$n$  = jumlah *term* yang untuk dalam dataset

$t$  = *term* dalam kalimat

## 2.6. Metode K-Means

*K-Means* merupakan salah satu metode *clustering* non hirarki yang bertujuan untuk membagi data yang ada menjadi satu atau lebih *cluster* atau kelompok, sehingga dapat mengelompokkan data dengan karakteristik yang sama ke dalam satu *cluster* dan menggabungkan data dengan karakteristik yang berbeda ke dalam *cluster* lain (Wahyudi, Masitha, Saragih, & Solikhun, 2020). Keunggulan utama dari algoritma *K-Means* adalah efisien, mudah diimplementasikan, umum digunakan dan sederhana. Kompleksitas waktu dari algoritma *K-Means* adalah  $O(tkn)$  dimana  $n$  sama dengan jumlah data, dimana  $k$  sama dengan jumlah *cluster*, dan  $t$  sama dengan jumlah iterasi. Jika  $k$  dan  $t$  lebih kecil dari  $n$ , maka algoritma *K-Means* merupakan algoritma yang memiliki hubungan linier dengan jumlah datanya (Liu, 2007 diacu dalam Robani & Widodo, 2016).

Langkah-langkah dengan algoritma *K-Means* yang dapat dilihat pada Gambar 2.2 sebagai berikut (Wanto, dkk., 2020).



**Gambar 2.2 Flowchart K-Means**

1. Tentukan jumlah *cluster*  $k$ .
2. Tentukan nilai pusat (*centroid*). Penentuan nilai *centroid* pada tahap awal dilakukan dengan cara *random*, sedangkan pada tahap iterasi digunakan rumus sebagai berikut:

$$V_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} X_{kj} \quad (4)$$

Keterangan :

$V_{ij}$  = *centroid* rata-rata *cluster* ke- $i$  untuk variabel ke- $j$

$N_i$  = jumlah anggota *cluster* ke- $i$

$i, k$  = indeks dari *cluster*

$j$  = indeks dari variabel

3. Pindahkan semua data atau objek ke *cluster* terdekat. Untuk mengukur jarak data ke pusat kelompok menggunakan jarak *Euclidean Distance* dengan rumus sebagai berikut:

$$D_{(i,j)} = \sqrt{(x_{1i} - x_{1j})^2 + \dots + (x_{ki} - x_{kj})^2} \quad (5)$$

Keterangan :

$D(i, j)$  = Jarak dari data ke- $i$  ke pusat *cluster*  $j$

$x_{ki}$  = Data ke- $i$  pada atribut data ke- $k$

$x_{kj}$  = Titik pusat ke- $j$  pada atribut ke- $k$

4. Kelompokkan objek berdasarkan jarak ke *centroid* terdekat.
5. Mengulangi langkah ke-3 dan langkah ke-4, lakukan iterasi hingga *centroid* bernilai optimal.

## 2.7. Metode K-Nearest Neighbors (K-NN)

*K-Nearest Neighbors* (K-NN) adalah salah satu metode teknik klasifikasi terhadap objek berlandaskan data pembelajaran yang jaraknya dekat dengan objek tersebut. *K-Nearest Neighbors* (K-NN) memiliki cara untuk menemukan kasus dengan menghitung kedekatan antara kasus baru dan kasus lama berdasarkan pencocokan bobot dari sejumlah fitur yang ada (Ndaumanu, Kursini, & Arief, 2014). *K-Nearest Neighbors* (K-NN) juga termasuk dalam algoritma *supervised learning*, dimana kelas dengan kemunculan terbanyak (mayoritas) akan menjadi kelas hasil dari klasifikasi (Mustakim & Oktaviani F, 2016). Sebelum menggunakan metode *K-Nearest Neighbors* (K-NN) untuk perhitungan, harus menentukan data latih dan data uji terlebih dahulu. Kemudian akan dilakukan proses perhitungan untuk mencari jarak menggunakan *Euclidean Distance*. Setelah itu, akan digunakan metode *K-Nearest Neighbors* (K-NN) untuk melakukan tahap perhitungan yang dapat dirumuskan sebagai berikut:

$$d_i = \sqrt{\sum_{i=0}^n (x_{2i} - x_{1i})^2} \quad (6)$$

Keterangan :

$x_1$  = sampel data

$x_2$  = data uji atau *testing*

$i$  = variabel data  
 $d$  = jarak  $d$   
 $n$  = dimensi data

Langkah-langkah algoritma *K-Nearest Neighbors* (K-NN) dapat dilihat pada Gambar 2.3 (Indrianto, Fauzi, & Muflikhah, 2017):



**Gambar 2.3** Flowchart *K-Nearest Neighbors*

1. Tentukan jumlah nilai  $k$  tetangga terdekat.
2. Hitung jarak masing – masing objek menggunakan *Euclidean distance*.
3. Urutkan hasil jarak berdasarkan nilai terkecil.
4. Mengambil data latih sejumlah  $k$  tetangga terdekat.
5. Menentukan kelas data baru berdasarkan mayoritas  $k$  tetangga terdekat.

Pada tahap ini, nilai hasil ekstraksi fitur dan kedekatan jarak antara data uji dan data latih, maka kalimat tersebut dipilih sebagai ringkasan. Posisi kalimat ringkasan akan berada pada urutan yang sama dengan kalimat asli dari dokumen (Desai & Shah, 2016). Keunggulan dari algoritma *K-Nearest Neighbors* (K-NN) adalah memiliki kemampuan terhadap data latih yang mempunyai banyak *noise* dan efektif ketika data latih besar (Yustanti, 2012).

## 2.8. ROUGE

*Recall Oriented Understudy for Gisting Evaluation (ROUGE)* merupakan evaluasi peringkasan teks yang digunakan untuk mengetahui kualitas hasil peringkasan yang mengevaluasi dan membandingkan hasil ringkasan sistem atau mesin dengan hasil ringkasan manual yang dibuat oleh manusia (*gold standard summary*). Salah satu jenis evaluasi peringkasan teks dalam *ROUGE* adalah *ROUGE-N*.

*ROUGE-N* pada dasarnya adalah perhitungan *recall* yang berdasarkan pada perbandingan *n-gram* antara hasil ringkasan manual yang dibuat oleh pakar dan hasil ringkasan yang dibuat oleh sistem atau mesin. Jumlah *n-gram* yang digunakan beragam, yaitu  $n=1-4$ . Namun, yang paling sering digunakan adalah *n-gram* dengan jumlah  $n=1$  atau *ROUGE-1* dan  $n=2$  atau *ROUGE-2* (Yuliska & Syaliman, 2020). *ROUGE-1* artinya pengukuran dengan konsep *unigram matching*, dimana pengukuran ini dihitung berdasarkan jumlah setiap satu kata (*unigram*) yang artinya mencari ketepatan berdasarkan *keyword* (kata – kata penting) yang dicari antara ringkasan sistem dengan ringkasan manual. Sedangkan, *ROUGE-2* artinya pengukuran dengan konsep *bigram matching*, dimana pengukuran ini dihitung berdasarkan jumlah setiap dua kata (*bigram*) yang artinya menghitung kemiripan antar kalimat antara ringkasan sistem dengan ringkasan manual. Persamaan *ROUGE-N* dapat dirumuskan sebagai berikut:

$$ROUGE - N = \frac{\sum_{S \in \text{SummRef}} \sum_{g \in \text{gram} \in S} \text{Count}_{\text{match}}(N\text{-gram})}{\sum_{S \in \text{SummRef}} \sum_{g \in \text{gram} \in S} \text{Count}(N\text{-gram})} \quad (7)$$

Keterangan:

$ROUGE-N$  = Hasil evaluasi peringkasan

$\sum_{S \in SummRef}$  = Jumlah kalimat referensi

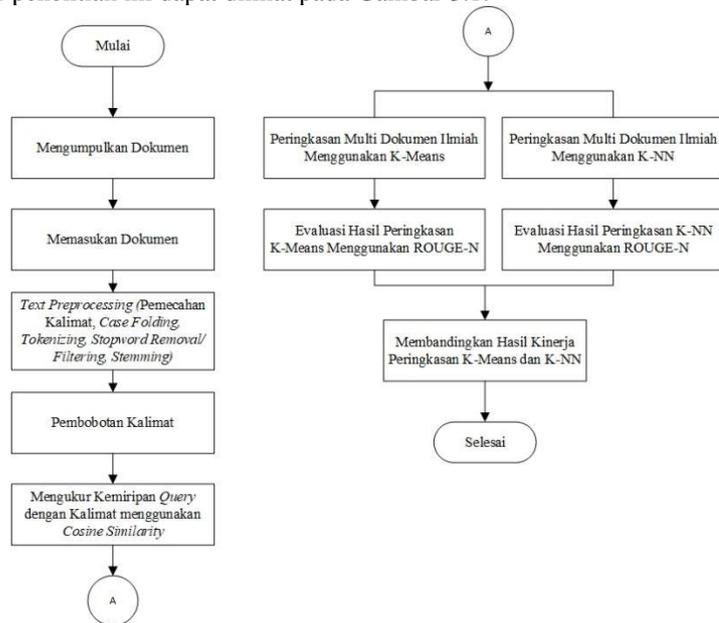
$\sum_{gramn \in S} Count_{match}(N - gram)$  = Jumlah  $N$ -gram yang sama pada ringkasan

$\sum_{gramn \in S} Count(N - gram)$  = Jumlah  $N$ -gram keseluruhan pada ringkasan

Evaluasi peringkasan menggunakan metode *Recall Oriented Understudy for Gisting Evaluation (ROUGE)* memiliki hasil perhitungan yang mempunyai nilai kesamaan 0 sampai 1. Nilai  $ROUGE-N$  yang semakin mendekati angka 1 dari setiap metode menunjukkan kualitas hasil ringkasan yang terbaik (Lin, 2004).

### 3. Metodologi

Alat yang digunakan dalam melaksanakan penelitian ini adalah sebuah laptop yang dilengkapi dengan beberapa *tools* yang digunakan seperti NLTK, Sastrawi, dan pyRouge. Bahan yang digunakan untuk penelitian ini adalah dokumen ilmiah berbahasa Indonesia sebanyak 30 dokumen dengan topik yang sama yaitu tentang “Sistem Informasi Perpustakaan”. Untuk memperlihatkan prosedur penelitian yang digunakan untuk mencapai tujuan penelitian. Diagram alir penelitian ini dapat dilihat pada Gambar 3.1.



**Gambar 3.1. Diagram Alir Penelitian**

Berdasarkan Gambar 3.1 ada beberapa proses yang dibutuhkan sebelum melakukan peringkasan menggunakan metode *K-Means* dan *K-Nearest Neighbors (K-NN)* diantaranya adalah sebagai berikut:

1. Mengumpulkan dokumen ilmiah sebanyak 30 dokumen dengan topik yang sama yaitu Sistem Informasi Perpustakaan. Dokumen yang dikumpulkan dalam bentuk elektronik dengan format .pdf kemudian dikonversi ke dalam bentuk teks dengan format .txt.
2. Ada beberapa aturan yang digunakan dalam memasukan dokumen ke dalam sistem, yaitu:
  - a. Judul pada dokumen akan dihilangkan.
  - b. Abstrak, nama penulis, dan instansi penulis akan dihilangkan
  - c. Judul bab dan sub bab akan dihilangkan.
  - d. Gambar dan tabel akan dihilangkan.
  - e. Daftar pustaka atau referensi akan dihilangkan.
3. Dokumen yang sudah dimasukkan ke dalam sistem akan dilakukan *text preprocessing*, yang terdiri dari beberapa tahap seperti Pemecahan Kalimat, *Case Folding*, *Tokenizing*, *Stopword Removal/Filtering*, dan *Stemming*.
4. Menghitung proses pembobotan kalimat atau *sentence scoring* menggunakan metode *Term Frequency – Inverse Sentence Frequency (TF-ISF)*.
5. Menghitung tingkat kemiripan *query* dengan kalimat menggunakan *Cosine Similarity*.
6. Proses peringkasan multi dokumen ilmiah berbahasa Indonesia menggunakan metode *K-Means* dilakukan dengan mengolah 10 dokumen, 20 dokumen, dan 30 dokumen dengan tingkat peringkasan sebanyak 5 kalimat, 10 kalimat, dan 15 kalimat dan menghasilkan keluaran berupa hasil ringkasan menggunakan metode *K-Means*. Pemilihan kalimat dilakukan dengan cara mengurutkan setiap *cluster* dari nilai tertinggi ke nilai

terendah yang diperoleh dari bobot nilai *Term Frequency – Inverse Sentence Frequency* (TF-ISF) dan mengambil nilai bobot kalimat tertinggi.

7. Proses peringkasan multi dokumen ilmiah berbahasa Indonesia menggunakan metode *K-Nearest Neighbors* (K-NN) dilakukan dengan mengolah 10, 20, dan 30 dokumen dengan tingkat peringkasan 5, 10, dan 15 kalimat dan menghasilkan keluaran berupa hasil ringkasan menggunakan metode *K-Nearest Neighbors* (K-NN). *K-Nearest Neighbors* (K-NN) akan menghitung jarak masing-masing objek dengan perhitungan jarak menggunakan *Cosine Similarity* yang telah dilakukan pada proses sebelumnya.
8. Hasil peringkasan multi dokumen ilmiah berbahasa Indonesia menggunakan metode *K-Means* dan *K-Nearest Neighbors* akan diuji dengan metode *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE). ROUGE mengevaluasi hasil ringkasan metode *K-Means* dan *K-Nearest Neighbors* (K-NN) dengan membandingkan hasil ringkasan sistem dan hasil ringkasan manual dari pakar. Pada penelitian ini digunakan metode ROUGE-N yaitu ROUGE-1 dan ROUGE-2 dari hasil peringkasan menggunakan *K-Means* dan *K-Nearest Neighbors* (K-NN) yang dapat dihitung dengan menggunakan Persamaan (7).
9. Membandingkan hasil kinerja ringkasan yang menggunakan *K-Means* dan *K-Nearest Neighbors* (K-NN) berdasarkan metode *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE). Nilai yang diperoleh dari evaluasi peringkasan teks pada metode ROUGE-N untuk hasil kinerja dari metode *K-Means* dan *K-Nearest Neighbors* adalah nilai presisi, *recall*, dan *f-measure* dalam melakukan peringkasan multi dokumen ilmiah berbahasa Indonesia. Untuk memberikan gambaran metode mana yang lebih baik dalam peringkasan multi dokumen ilmiah berbahasa Indonesia.

#### 4. Hasil dan Analisis

Hasil peringkasan multi dokumen ilmiah berbahasa Indonesia menggunakan metode *K-Means* dan *K-Nearest Neighbors* (K-NN) dievaluasi ringkasannya menggunakan metode *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE), yakni ROUGE-1 dan ROUGE-2. Evaluasi peringkasan ini berguna untuk membandingkan hasil peringkasan menggunakan sistem dan hasil peringkasan manual dari pakar atau ahli. Dalam pengujian ini diperoleh nilai presisi, *recall*, dan *f-measure*. Hasil kinerja dari metode *K-Means* dan *K-Nearest Neighbors* (K-NN) dapat dilihat pada Tabel 4.1.

**Tabel 4.1 Hasil Pengujian ROUGE-1 dan ROUGE-2 Peringkasan Multi Dokumen Ilmiah dari Metode K-Means dan K-Nearest Neighbors (K-NN)**

Jumlah Dokumen	Tingkat Peringkasan	ROUGE-1						ROUGE-2					
		K-Means			K-Nearest Neighbors			K-Means			K-Nearest Neighbors		
		Presisi	Recall	F-Measure	Presisi	Recall	F-Measure	Presisi	Recall	F-Measure	Presisi	Recall	F-Measure
10 Dokumen	5 kalimat	0.235	0.139	0.173	0.307	0.165	0.215	0.035	0.020	0.026	0.012	0.006	0.009
	10 kalimat	0.248	0.221	0.234	0.31	0.221	0.258	0.023	0.020	0.021	0.023	0.016	0.019
	15 kalimat	0.302	0.225	0.258	0.362	0.296	0.326	0.029	0.021	0.025	0.045	0.037	0.045
20 Dokumen	5 kalimat	0.25	0.113	0.156	0.217	0.158	0.183	0.00	0.00	0.00	0.008	0.006	0.007
	10 kalimat	0.234	0.168	0.195	0.237	0.198	0.216	0.00	0.00	0.00	0.012	0.010	0.011
	15 kalimat	0.243	0.261	0.252	0.253	0.25	0.251	0.028	0.030	0.029	0.015	0.015	0.015
30 Dokumen	5 kalimat	0.145	0.059	0.084	0.217	0.158	0.183	0.00	0.00	0.00	0.008	0.006	0.007
	10 kalimat	0.321	0.152	0.207	0.233	0.204	0.217	0.018	0.008	0.011	0.011	0.010	0.010
	15 kalimat	0.391	0.277	0.324	0.236	0.261	0.248	0.067	0.048	0.056	0.013	0.015	0.014

Dari penelitian yang dilakukan dengan menguji kinerja dari masing – masing metode menggunakan sejumlah dokumen yang telah ditentukan, maka didapatkan hasil kinerja yang diperoleh dengan nilai presisi, *recall*, dan *f-measure* yang dapat dilihat pada Tabel 4.2. Nilai yang dibandingkan dari hasil evaluasi peringkasan menggunakan metode ROUGE-N adalah nilai *f-measure*, dimana nilai *f-measure* tertinggi digunakan untuk menyatakan kualitas ringkasan sistem dan ringkasan manual. Nilai ROUGE-N memiliki nilai antara 0 sampai 1 dimana jika hasil nilai = 0, maka hasil ringkasan sistem tidak mirip sama sekali dengan hasil peringkasan ahli atau manual. Sedangkan, jika hasil nilai = 1, maka hasil ringkasan sistem identik dengan hasil peringkasan ahli atau manual.

**Tabel 4.2 Perbandingan Hasil Kinerja Peringkasan Multi Dokumen Ilmiah Menggunakan Metode K-Means dan K-Nearest Neighbors (K-NN)**

Metode			ROUGE-1	ROUGE-2
--------	--	--	---------	---------

	Jumlah Dokumen	Tingkat Peringkasan	Presisi	Recall	F-Measure	Presisi	Recall	F-Measure
<i>K-Means</i>	30 Dokumen	15 Kalimat	0.391	0.277	0.324	0.067	0.048	0.056
<i>K-Nearest Neighbors</i>	10 Dokumen	15 Kalimat	0.362	0.296	0.326	0.045	0.037	0.045

Tabel 4.2 menunjukkan bahwa pada *ROUGE-1* hasil optimal dari metode *K-Means* diperoleh dengan nilai presisi 0.391, *recall* 0.277, dan *f-measure* 0.324 dengan jumlah 30 dokumen dan tingkat peringkasan sebanyak 15 kalimat. Selanjutnya pada *ROUGE-2* dari metode *K-Means*, hasil yang optimal diperoleh dengan nilai presisi 0.067, *recall* 0.048, dan *f-measure* 0.056 dengan jumlah 30 dokumen dan tingkat peringkasan sebanyak 15 kalimat. Sedangkan pada *ROUGE-1* dari metode *K-Nearest Neighbors* (K-NN), hasil yang paling optimal diperoleh dengan nilai presisi 0.362, *recall* 0.296, dan *f-measure* 0.326 dengan jumlah 10 dokumen dan tingkat peringkasan sebanyak 15 kalimat. Kemudian pada *ROUGE-2* dari metode *K-Nearest Neighbors* (K-NN) hasil optimal diperoleh dengan nilai presisi 0.045, *recall* 0.037, dan *f-measure* 0.045 dengan jumlah 10 dokumen dan tingkat peringkasan sebanyak 15 kalimat.

Berdasarkan pengujian dan analisis perbandingan hasil dari metode *K-Means* dan *K-Nearest Neighbors* (K-NN) yang telah dilakukan, dapat disimpulkan bahwa kedua metode memiliki karakteristik dan keunggulan masing – masing. Metode *K-Means* memiliki kinerja yang lebih baik jika dokumen yang digunakan relatif lebih banyak. Sebaliknya, metode *K-Nearest Neighbors* (K-NN) memiliki kinerja yang lebih baik jika dokumen yang digunakan relatif lebih sedikit.

## 5. Kesimpulan dan Saran

Hasil kinerja yang didapatkan pada penelitian adalah metode *K-Means* memiliki kinerja yang lebih baik jika dokumen yang digunakan relatif lebih banyak. Sebaliknya, metode *K-Nearest Neighbors* (K-NN) memiliki kinerja yang lebih baik jika dokumen yang digunakan relatif lebih sedikit. Pada *ROUGE-1* hasil optimal pada metode *K-Means* diperoleh dengan nilai presisi 0.391, *recall* 0.277, dan *f-measure* 0.324 untuk 30 dokumen dan tingkat peringkasan sebanyak 15 kalimat. Sedangkan pada metode *K-Nearest Neighbors* (K-NN), hasil yang paling optimal diperoleh dengan nilai presisi 0.362, *recall* 0.296, dan *f-measure* 0.326 untuk 10 dokumen dan tingkat peringkasan sebanyak 15 kalimat. Selanjutnya pada *ROUGE-2* pada metode *K-Means*, hasil yang optimal diperoleh dengan nilai presisi 0.067, *recall* 0.048, dan *f-measure* 0.056 untuk 30 dokumen dan tingkat peringkasan sebanyak 15 kalimat. Sedangkan pada metode *K-Nearest Neighbors* (K-NN), hasil yang paling optimal diperoleh dengan nilai presisi 0.045, *recall* 0.037 dan *f-measure* 0.045 untuk 10 dokumen dan tingkat peringkasan sebanyak 15 kalimat.

Untuk penelitian dan analisis lebih lanjut mengenai peringkasan multi dokumen ilmiah berbahasa Indonesia, disarankan untuk melakukan hal sebagai berikut:

1. Dalam pengambilan topik untuk peringkasan multi dokumen ilmiah, sebaiknya mencari topik yang tidak terlalu banyak mengandung gambar, tabel, dan rumus.
2. Perlu dilakukan penelitian lebih lanjut menggunakan algoritma lain untuk mengetahui perbandingan algoritma tersebut sehingga hasil penelitian yang didapat lebih beragam.
3. Untuk *stopword removal* dan *stemming* dalam Bahasa Indonesia pada *tools* NTLK dan *library* Sastrawi, sebaiknya menambahkan kondisi kata (*term*) lain pada *library* yang digunakan agar *corpus* yang dihasilkan lebih baik.
4. Untuk proses peringkasan manual sebaiknya mencari pakar seperti Dosen atau Guru Bahasa Indonesia yang ahli di bidang tersebut.

## Daftar Pustaka:

- Abdullah, M. Z., & Fatchah, C. (2018). Peringkasan Multi – Dokumen Berita Berdasarkan Fitur Berita dan Part of Speech Tagging. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, 4(2), 56–66.
- Anam, C. (2019). Meyiapkan Dokumen Ilmiah untuk Publikasi di Jurnal. *Berkala Fisika*, 22(3), 86–88.
- Aysiyah, N., & Inawati. (2020). Perilaku Penelusuran Informasi Mahasiswa D3. *BIBLIOTIKA: Jurnal Kajian Perpustakaan dan Informasi*, 4(2), 229–237
- Desai, N., & Shah, P. (2016). Automatic Text Summarization using Supervised Machine Learning Technique for Hindi Language. *IJRET: International Journal of Research in Engineering and Technology*, 05(06), 361–367.
- Husni, Negara, Y. D. P., & Syarief, M. (2015). Clusterisasi Dokumen Web (Berita) Bahasa Indonesia Menggunakan Algoritma K-Means. *Jurnal Simantec*, 4(3), 159–166.

- Indrianto, R., Fauzi, M. A., & Muflikhah, L. (2017). Peringkasan Teks Otomatis Pada Artikel Berita Kesehatan Menggunakan K-Nearest Neighbor Berbasis Fitur Statistik. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIHK) Universitas Brawijaya*, 1(11), 1–6.
- Jumeilah, F. S. (2017). Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 1(1), 19–25.
- Lin, C. Y. (2004). *ROUGE: a Package for Automatic Evaluation of Summaries*. [Prosiding] Workshop on Text Summarization Brances Out, Barcelona: 25 – 26 July 2004, 74-81.
- Mustakim, & Oktaviani F, G. (2016). Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa. *Jurnal Sains, Teknologi dan Industri*, 13(2), 195–202.
- Mustaqhfiri, M., Abidin, Z., & Kusumawati, R. (2011). Peringkasan Teks Otomatis Berita Berbahasa Indonesia Menggunakan Metode Maximum Marginal Relevance. *Jurnal Ilmu Komputer dan Teknologi Informasi*, 4(4), 134–147.
- Ndaumanu, R. I., & Arief, Kusrini, M. R. (2014). Analisis Prediksi Tingkat Pengunduran Diri Mahasiswa dengan Metode K-Nearest Neighbor. *Jatsi*, 1(1), 1–15.
- Pradnyana, G. A., & Mogi, I. K. A. (2014). Implementasi Automated Text Summarization untuk Dokumen Tunggal Berbahasa Indonesia dengan Menggunakan Graph-Based. *Jurnal Ilmiah NERO*, 1(2), 33–46.
- Purnamasari, K. K., & Widiastuti, N. I. (2017). Perbandingan Algoritma K-Means dan K-Nearest Neighbors pada Sistem Peringkasan Otomatis. *Jurnal Ilmiah Komputer dan Informatika*, 6(2), 57–66.
- Rautray, R., & Balabantaray, R. C. (2018). An Evolutionary Framework for Multi Document Summarization Using Cuckoo Search Approach: MDSCSA. *Applied Computing and Informatics*, 14(2), 134–144.
- Robani, M., & Widodo, A. (2016). Algoritma K-Means Clustering untuk Pengelompokan Ayat Al Quran pada Terjemahan Bahasa Indonesia. *Jurnal Sistem Informasi Bisnis*, 6(2), 164–176.
- Saraswati, N. F., Indriati, & Perdana, R. S. (2018). Peringkasan Teks Otomatis Menggunakan Metode Maximum Marginal Relevance pada Hasil Pencarian Sistem Temu Kembali Informasi untuk Artikel Berbahasa Indonesia. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIHK) Universitas Brawijaya*, 2(11), 5494–5502.
- Tardan, P. P., Erwin, A., Eng, K. I., & Muliady, W. (2013). Automatic Text Summarization Based on Semantic Analysis Approach for Documents in Indonesian Language. *2013 International Conference on Information Technology and Electrical Engineering (ICITEE)*, 47–52.
- Wahyudi, Masitha, Saragih, & Solikhun. (2020). *Data Mining: Penerapan Algoritma K-Means Clustering dan K-Medoids Clustering* (J. Simarmata (ed.)). Yayasan Kita Menulis.
- Wanto, A., et al. (2020). *Data Mining: Algoritma dan Implementasi* (T. Limbong (ed.)). Yayasan Kita Menulis.
- Yuliska, Y., & Syaliman, K. U. (2020). Literatur Review Terhadap Metode, Aplikasi dan Dataset Peringkasan Dokumen Teks Otomatis untuk Teks Berbahasa Indonesia. *IT Journal Research and Development (ITJRD)*, 5(1), 19–31.
- Yustanti, W. (2012). Algoritma K-Nearest Neighbour untuk Memprediksi Harga Jual Tanah. *Jurnal Matematika Statistika dan Komputasi*, 9(1), 57–68.