ANALISIS PERBANDINGAN *K-MEANS CLUSTERING* DAN *TEXTRANK* DALAM PERINGKASAN TEKS BERITA BERBAHASA INDONESIA

SKRIPSI

Oleh:
MUHAMMAD DAFFA PRAMUDITYA SAPUTRA
NIM. 210605110010

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025

ANALISIS PERBANDINGAN *K-MEANS CLUSTERING* DAN *TEXTRANK* DALAM PERINGKASAN TEKS BERITA BERBAHASA INDONESIA

SKRIPSI

Diajukan kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

Oleh :
MUHAMMAD DAFFA PRAMUDITYA SAPUTRA
NIM. 210605110010

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025

HALAMAN PERSETUJUAN

ANALISIS PERBANDINGAN *K-MEANS CLUSTERING* DAN *TEXTRANK* DALAM PERINGKASAN TEKS BERITA BERBAHASA INDONESIA

SKRIPSI

Oleh :
MUHAMMAD DAFFA PRAMUDITYA SAPUTRA
NIM. 210605110010

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 21 Maret 2025

Pembimbing I,	Pembimbing II,
Prof. Dr. Muhammad Faisal, M.T	Dr. M. Imamudin, Lc., MA
NIP. 19740510 200501 1 007	NIP. 19740602 200901 1 010

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Ir. Fachrul Kurniawan, M.MT., IPU
NIP. 19771020 200912 1 001

HALAMAN PENGESAHAN

ANALISIS PERBANDINGAN *K-MEANS CLUSTERING* DAN *TEXTRANK* DALAM PERINGKASAN TEKS BERITA BERBAHASA INDONESIA

SKRIPSI

Oleh :
MUHAMMAD DAFFA PRAMUDITYA SAPUTRA
NIM. 210605110010

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 15 Mei 2025

Susunan Dewan Penguji

Ketua Penguji	: Dr. Zainal Abidin, M.Kom NIP. 19760613 200501 1 004	( )
Anggota Penguji I	: Hani Nurhayati, M.T NIP. 19780625 200801 2 006	( )
Anggota Penguji II	: Prof. Dr. Muhammad Faisal, M.T NIP. 19740510 200501 1 007	( )
Anggota Penguji III	: Dr. M. Imamudin, Lc., MA NIP. 19740602 200901 1 010	( )

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Ir. Fachrul Kurniawan, M.MT., IPU
NIP. 19771020 200912 1 001

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	: Muhammad Daffa Pramuditya Saputra
NIM	: 210605110010
Fakultas / Program Studi	: Sains dan Teknologi / Teknik Informatika
Judul Skripsi	: Analisis Perbandingan *K-Means Clustering* dan *Textrank* Dalam Peringkasan Teks Berita Berbahasa Indonesia

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 16 Mei 2025
Yang membuat pernyataan,

Muhammad Daffa Pramuditya Saputra
NIM.210605110010

MOTTO

"The sun rises everyday, no matter how dark the night was"

HALAMAN PERSEMBAHAN

Segala puji syukur dipanjatkan ke hadirat Tuhan Yang Maha Esa, atas anugerah tak terhingga berupa kekuatan lahir dan batin, pencerahan akal budi, serta ketabahan jiwa, sehingga karya ilmiah ini dapat dirampungkan sebagai bukti dedikasi dalam menuntut ilmu. Dengan segala kerendahan hati dan penghormatan yang mendalam, karya ini kupersembahkan kepada:

Kedua Mutiara Hati,

Mama Marviyani Candrasari dan Papa Hartono

Sang penjaga api kehidupan yang tak pernah redup, pelita dalam gelap gulita perjalanan hidup, yang dengan tangan penuh kasih telah mengukir jalan menuju masa depan.

Teruntuk Oma tersayang, Truna Dahrita

Terima kasih atas kebijaksanaan yang menerangi, kasih sayang yang tak pernah surut, serta menjadi sumber kekuatan dan keteduhan dalam setiap langkah perjalanan hidup penulis.

Para Pejuang Ilmu Pengetahuan,

Seluruh insan akademis Teknik Informatika Angkatan 2021, yang telah bersama-sama mengarungi samudra pengetahuan. Semoga persaudaraan intelektual ini abadi melampaui ruang dan waktu, membawa kita semua pada puncak pencapaian yang indah.

KATA PENGANTAR

Bismillahirrahmaanirrahiim,

Penulis memanjatkan puji syukur ke hadirat Allah SWT atas limpahan rahmat dan hidayah-Nya sehingga penyusunan skripsi dengan judul "Analisis Perbandingan *K-Means Clustering* dan *Textrank* Dalam Peringkasan Teks Berita Berbahasa Indonesia" dapat terselesaikan dengan baik. Skripsi ini ditulis untuk memenuhi salah satu syarat dalam meraih gelar sarjana pada Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang.

Penulis menyadari sepenuhnya bahwa penyelesaian karya ilmiah ini dapat terlaksana berkat bantuan, arahan, dan dukungan dari berbagai pihak. Atas dasar itu, penulis mengucapkan terima kasih yang tulus kepada:

1. Prof. Dr. M. Zainuddin, M.A., selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang, atas langkah-langkah progresif dalam peningkatan fasilitas dan infrastruktur akademik, yang telah berkontribusi pada pencapaian standar pendidikan yang unggul.
2. Prof. Dr. Hj. Sri Harini, M.Si., selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang, atas dedikasi dan komitmennya dalam meningkatkan mutu akademik di lingkungan fakultas.
3. Dr. Ir. Fachrul Kurniawan, M.MT., IPU., selaku Ketua Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim

Malang, atas bimbingan dan dukungan yang telah diberikan sepanjang masa studi.

4. Prof. Dr. Muhammad Faisal, M.T., selaku dosen pembimbing 1, yang dengan penuh dedikasi telah memberikan arahan, masukan konstruktif, serta motivasi yang sangat berharga bagi penulis sepanjang pelaksanaan penelitian.
5. Dr. M. Imamudin Lc, MA., selaku dosen pembimbing 2, atas ketelitian dan panduan yang telah memberikan kontribusi signifikan dalam penyempurnaan penelitian ini.
6. Dr. Zainal Abidin, M.Kom., dan Hani Nurhayati, M.T., selaku dosen penguji, atas masukan kritis dan rekomendasi yang sangat bermanfaat dalam meningkatkan kualitas penelitian ini.
7. Fatchurrohman, M.Kom., selaku dosen wali, atas kepedulian dan pendampingan yang telah diberikan sepanjang perjalanan akademik penulis.
8. Khadijah Fahmi Hayati Holle, M.Kom dan Tri Mukti Lestari, M.Kom yang telah memberikan pemahaman mendalam mengenai *Natural Language Processing* dan *Information Retrieval* sehingga menginspirasi penulis untuk mengembangkan penelitian pada ranah tersebut.
9. Segenap dosen dan tenaga kependidikan Program Studi Teknik Informatika, atas dedikasi dalam mentransfer ilmu pengetahuan serta penyediaan sarana pembelajaran yang mendukung selama periode perkuliahan.

10. Mama Marviyani Candrasari, Papa Hartono, serta Oma Truna Dahrita, atas kasih sayang yang melimpah, untaian doa yang tulus, dan dukungan moral yang tak pernah putus. Nasihat bijak dan perhatian yang diberikan telah menjadi sumber inspirasi dan motivasi utama penulis dalam merampungkan pendidikan ini.
11. Annisa Fitri Madani, selaku rekan diskusi dan partner yang andal, atas dukungan moral dan motivasi yang konsisten dalam setiap fase perjalanan akademik, baik dalam situasi yang menggembirakan maupun penuh tantangan. Kehadiran dan dorongan semangat yang diberikan telah menjadi kekuatan tersendiri bagi penulis dalam menyelesaikan studi ini.
12. Segenap rekan Teknik Informatika khususnya Angkatan 2021 "ASTER", atas kontribusi pengetahuan, motivasi, dan kenangan tak terlupakan yang telah diberikan. Semoga ikatan persaudaraan kita terus terjaga dan kesuksesan menyertai langkah kita masing-masing.

Penulis memahami sepenuhnya bahwa karya ilmiah ini masih memerlukan penyempurnaan. Berbagai masukan dan saran yang konstruktif akan menjadi kontribusi berharga untuk pengembangan penelitian ini. Harapan penulis, semoga karya ini dapat berkontribusi pada kemajuan keilmuan, terutama dalam konteks integrasi nilai-nilai keislaman.

Malang, 25 Mei 2025

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Saputra, Muhammad Daffa Pramuditya. 2025. **Analisis Perbandingan K-Means Clustering dan Textrank dalam Peringkasan Teks Berita Berbahasa Indonesia.** Skripsi. Program Studi Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Prof. Dr. Muhammad Faisal, M.T (II) Dr. M. Imamudin Lc, MA

Kata Kunci: Peringkasan Teks Otomatis, *K-Means Clustering*, *TextRank*

Penelitian ini membandingkan kinerja metode *K-Means Clustering* dan *Textrank* dalam peringkasan teks otomatis berita berbahasa Indonesia. Menggunakan dataset IndoSum dengan 2.500 dokumen berita, kedua metode dievaluasi pada lima tingkat kompresi berbeda (30%, 40%, 50%, 60%, dan 70%) dengan metrik ROUGE-1, ROUGE-2, dan ROUGE-L. Hasil menunjukkan bahwa metode *Textrank* mengungguli *K-Means Clustering* pada tingkat kompresi rendah hingga menengah, dengan nilai *F1-score* tertinggi 0.589 pada kompresi 30%, sementara *K-Means Clustering* menunjukkan performa yang lebih stabil pada tingkat kompresi tinggi yaitu 60-70%. Penelitian ini juga mengidentifikasi fenomena menarik di mana nilai *recall* meningkat sementara *precision* menurun seiring bertambahnya tingkat kompresi. Untuk ROUGE-2, *Textrank* mencapai performa terbaik pada kompresi 40% dengan nilai *F1-score* 0.501, sedangkan untuk ROUGE-L pada kompresi yang sama dengan nilai 0.531. Keseimbangan optimal antara cakupan informasi dan ketepatan ditemukan pada rentang kompresi 30-50%. Hasil ini memberikan landasan untuk pemilihan metode peringkasan yang optimal sesuai dengan kebutuhan ringkasan pada berbagai tingkat kompresi untuk dokumen berita berbahasa Indonesia.

ABSTRACT

Saputra, Muhammad Daffa Pramuditya. 2025. **Comparative Analysis of K-Means Clustering and TextRank for Indonesian News Text Summarization.** Thesis. Informatics Engineering Study Program, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University of Malang. Supervisor: (I) Prof. Dr. Muhammad Faisal, M.T (II) Dr. M. Imamudin Lc, MA

Keywords: Automatic Text Summarization, K-Means Clustering, TextRank

This research compares the performance of K-Means Clustering and TextRank methods in automatic text summarization of Indonesian news. Using the IndoSum dataset with 2,500 news documents, both methods were evaluated at five different compression levels (30%, 40%, 50%, 60%, and 70%) using ROUGE-1, ROUGE-2, and ROUGE-L metrics. Results show that the TextRank method outperforms K-Means Clustering at low to medium compression levels, with the highest F1-score of 0.589 at 30% compression, while K-Means Clustering demonstrates more stable performance at higher compression levels (60-70%). The research also identifies an interesting phenomenon where recall values increase while precision decreases as compression levels rise. For ROUGE-2, TextRank achieves best performance at 40% compression with an F1-score of 0.501, while for ROUGE-L at the same compression level with a value of 0.531. The optimal balance between information coverage and accuracy was found in the 30-50% compression range. These findings provide an empirical basis for selecting the optimal summarization method according to summary requirements at various compression levels for Indonesian news documents.

الملخص

سابوترا، محمد دفا براموديتيا. 2025. التحليل المقارن لتجميع في تلخيص النصوص الإخبارية الإندونيسية. رسالة جامعية. برنامج دراسة هندسة المعلوماتية، كلية العلوم والتكنولوجيا، الجامعة الإسلامية الحكومية، مولانا مالك إبراهيم مالانج. المشرف: (أولاً) الدكتور محمد فيصل، م.ت (ثانيًا) الدكتور إمام الدين ، ماجستير

الكلمات الرئيسية: التلخيص التلقائي للنصوص، التلخيص التلقائي للنصوص، التجميع العنقودي ك-مجموعة وسائل ,تيكسترك

يقارن هذا البحث بين أداء طريقتَي التجميع العنقودي ك-مجموعة وسائل ,تيكسترك في التلخيص التلقائي للنصوص التي تحتوي على 2,500 مستند إخباري، تم تقييم كلتا الطريقتين على IndoSum للأخبار الإندونيسية. باستخدام مجموعة بيانات التجميع العنقودي ، ROUGE-1 خمسة مستويات ضغط مختلفة (30% ,40% ,50% ,60% ,70%) باستخدام مقاييس تُظهر النتائج أن طريقة تيكسترك تتفوق على طريقة عند مستويات ضغط .ROUGE-Lو ROUGE-2, ك-مجموعة وسائل عند ضغط 30%، بينما تُظهر طريقة التجميع العنقودي ك- 0.589 F1 منخفضة إلى متوسطة، حيث بلغت أعلى قيمة لنتيجة مجموعة وسائل أداءً أكثر استقرارًا عند مستويات ضغط عالية تتراوح بين ,70%,60% حدد هذا البحث أيضًا ظاهرة مثيرة للاهتمام حقق ,تيكسترك أفضل أداء عند ،ROUGE-2 حيث تزداد قيمة الاسترجاع بينما تقل الدقة مع زيادة مستوى الضغط. بالنسبة إلى عند نفس الضغط 0.531. تم العثور على التوازن ROUGE-L تبلغ 0.501، بينما حقق F1 ضغط بنسبة 70% مع درجة الأمثل بين تغطية المعلومات والدقة في نطاق ضغط يتراوح بين% 30-50%. توفر هذه النتائج أساسًا لاختيار طريقة التلخيص المثلى .وفقًا لاحتياجات التلخيص عند مستويات ضغط مختلفة للوثائق الإخبارية الإندونيسية

BAB I

PENDAHULUAN

1.1 Latar Belakang

Era komunikasi dan teknologi informasi telah berkembang pesat dan berpengaruh secara signifikan pada kegiatan sehari-hari, terutama dalam akses berita dan informasi. Berita, yang terdiri dari kumpulan teks informasi mengenai peristiwa terkini, kini dapat disebarluaskan melalui berbagai platform. Di era digital, informasi dan berita dapat diakses dengan cepat, menyebabkan peningkatan eksponensial dalam volume teks. Informasi ini hadir dalam berbagai bentuk, seperti artikel berita, laporan penelitian, dokumen, hingga tulisan di media sosial (Ronaning Roem & Vanisya, 2024).

Pada dasarnya, manusia memiliki kebutuhan mendasar untuk mendapatkan informasi. Informasi membantu individu dalam berbagai hal, mulai dari pengambilan keputusan, penyelesaian masalah, hingga pengembangan diri (Lubis & Koto, 2020). Informasi yang akurat dan relevan akan mendorong seseorang untuk mengambil keputusan agar lebih bijak, baik dalam aspek personal, profesional, maupun sosial. Oleh karena itu, akses terhadap informasi dan berita berkualitas menjadi salah satu faktor penting bagi masyarakat secara keseluruhan.

Pemanfaatan teknologi informasi yang baik dan bijak diharapkan dapat membantu masyarakat memenuhi kebutuhan informasi. Penggunaan teknologi informasi mempermudah pengolahan, analisis, dan komunikasi, sehingga pengelolaan informasi menjadi lebih efisien (Robiyanto et al., 2019). Saat ini

berbagai bentuk informasi, seperti berita, jurnal, dan artikel ilmiah, sudah dapat diakses melalui internet dan media sosial. Seiring perkembangan teknologi, jumlah berita berbahasa Indonesia yang diproduksi oleh media massa dan platform digital semakin meningkat setiap tahunnya, hal ini mengakibatkan berpengaruhnya penyebaran informasi dan tanggapan publik di masyarakat.

Informasi dalam berita sering kali beragam dan padat, sehingga pembaca mungkin kesulitan menangkap inti dari informasi yang disampaikan. Untuk mengatasi masalah ini, ringkasan berita dapat menyajikan informasi penting secara singkat dan mudah dipahami, tanpa perlu membaca keseluruhan artikel. Bagi pembaca yang memiliki kesibukan tinggi dan waktu terbatas, ringkasan berita memungkinkan mereka mendapatkan informasi secara cepat dan tepat, membantu pembaca memanfaatkan waktu secara lebih efisien dalam mengikuti perkembangan informasi terkini.

Oleh karena itu, diperlukan peringkas teks otomatis yang mampu menyaring dan meringkas informasi penting secara relevan dan singkat. Peringkas teks otomatis (*Automatic Text Summarization*) menyajikan informasi inti dari suatu dokumen atau artikel berita dengan hasil yang biasanya tidak lebih dari setengah panjang dokumen aslinya (Suputra, 2017). Secara umum, peringkasan teks otomatis mempunyai dua pendekatan metode yaitu ekstraktif dan abstraktif. Metode ekstraktif menghasilkan kalimat penting dari dokumen asli tanpa merubah kata, sedangkan metode abstraktif menghasilkan hasil seperti halnya parafrase, dengan kalimat-kalimat yang mendekati hasil ringkasan manusia (Khatri et al., 2018).

Berbagai penelitian telah dilakukan dalam bidang peringkasan teks otomatis. (Khan et al., 2019) memperkenalkan pendekatan ekstraktif menggunakan *K-Means Clustering* dan TF-IDF untuk peringkasan teks, dimana kalimat-kalimat dikelompokkan berdasarkan kesamaan semantik dan dipilih berdasarkan representasi vektor TF-IDF mereka. Penelitian tersebut menyatakan bahwa *K-Means Clustering* efektif dalam mengidentifikasi kluster kalimat yang memiliki topik serupa. Sementara itu, penelitian (Husniah et al., 2022) membahas tentang peringkasan teks otomatis menggunakan artikel berbahasa Indonesia menggunakan metode *TextRank*, memperoleh nilai ROUGE-1 sebesar 68.76% pada *compression rate* 50%. Penelitian tersebut menunjukkan bahwa algoritma *TextRank* efektif dalam menentukan kalimat-kalimat penting dalam dokumen berdasarkan graf kemiripan antar kalimat.

Penelitian yang dilakukan (Shetty & Kallimani, 2017) mengimplementasikan metode *K-Means Clustering* untuk peringkasan teks ekstraktif, di mana kalimat direpresentasikan sebagai vektor dalam ruang dimensi berdasarkan *vocabulary* dokumen. Metode ini berhasil mengelompokkan kalimat berdasarkan tingkat kesamaan semantik, memungkinkan pemilihan kalimat representatif dari setiap *cluster* untuk membentuk ringkasan akhir. Perlu dicatat bahwa dalam penelitian tersebut, jumlah *cluster* yang dibentuk ditentukan berdasarkan tingkat kompresi yang diinginkan. Shetty dan Kallimani mengatur tingkat kompresi sebesar 30%, artinya ringkasan akhir berisi sekitar 30% dari jumlah kalimat dalam dokumen asli. Pendekatan ini memastikan bahwa ringkasan yang dihasilkan mencakup informasi penting dari berbagai topik dalam dokumen.

Berdasarkan penelitian-penelitian tersebut, penulis memilih untuk menganalisis perbandingan implementasi antara metode *K-Means Clustering* dan *TextRank* pada peringkasan teks berita berbahasa Indonesia. Dalam analisis perbandingan ini, *K-Means Clustering* dievaluasi kemampuannya dalam menggabungkan kalimat-kalimat berdasarkan kemiripan semantiknya, sementara *TextRank* dinilai dari segi efektivitasnya dalam mengurutkan kalimat-kalimat berdasarkan tingkat kepentingannya.

K-Means Clustering adalah teknik pengelompokan data yang memisahkan dataset menjadi beberapa kelompok berbeda, di mana setiap kelompok memiliki titik pusat atau *centroid* tersendiri sebagai representasi karakteristik kelompok tersebut (Yuan & Yang, 2019). Tujuan dari *K-Means* yaitu untuk memaksimalkan variasi antar kelompok dan suatu kelompok secara keseluruhan. Sementara itu, *TextRank* merupakan metode berdasarkan graf dimana setiap kalimat yang ada dalam dokumen digambarkan sebagai simpul dan hubungan antara kalimat sebagai sisi (Andriani et al., 2019). Analisis perbandingan kedua metode ini perlu dilakukan untuk memahami manfaat dan kelemahan dari masing-masing metode untuk menghasilkan ringkasan berita berbahasa Indonesia.

Beberapa penelitian sebelumnya telah mempertimbangkan tingkat kompresi dalam evaluasi metode peringkasan teks. (Shetty & Kallimani, 2017) menerapkan *K-Means Clustering* dengan tingkat kompresi 30% dan mengevaluasi hasilnya menggunakan *precision* dan *recall*. Sementara itu, (Husniah et al., 2022) mengimplementasikan algoritma *TextRank* untuk artikel berbahasa Indonesia dan mengevaluasi kinerja metode tersebut pada tingkat kompresi yang berbeda (30%

dan 50%) dengan metrik ROUGE-1, ROUGE-2, dan ROUGE-L digunakan. Berdasarkan beberapa penelitian yang sudah dipaparkan, penulis tertarik untuk mengeksplorasi lebih lanjut bagaimana kinerja *K-Means Clustering* dibandingkan dengan *TextRank* ketika diterapkan pada dokumen berbahasa Indonesia, serta bagaimana performa kedua metode tersebut pada beberapa tingkat kompresi yang berbeda.

Sebagai umat Muslim, penting untuk memahami bahwa setiap ilmu dan teknologi yang dikembangkan harus berlandaskan pada prinsip-prinsip yang bermanfaat bagi seluruh umat manusia. Pengembangan sistem peringkas teks otomatis tidak hanya merupakan kemajuan teknologi, tetapi juga upaya untuk memberikan informasi yang berkualitas tinggi dengan singkat, yang membantu orang untuk memahami arti dan makna dari sebuah informasi. Hal tersebut sejalan dengan sabda Rasulullah SAW yang berbunyi:

بُعِثْتُ بِجَوَامِعِ الكَلِمِ

"Aku diutus dengan Al Jawami'ul Kalim" (HR. Bukhari)

Hal tersebut menandakan bahwa Rasulullah diberikan kemampuan untuk memberikan kata-kata yang singkat tetapi mendalam (Alhadi, 2022). Prinsip ini mencerminkan pentingnya menyampaikan informasi yang padat, ringkas, dan bermanfaat bagi orang lain, sebagaimana peringkasan teks bertujuan untuk memberikan intisari informasi tanpa mengurangi esensinya. Dalam konteks ini, teknologi juga harus berperan dalam membawa manfaat bagi masyarakat,

sebagaimana ajaran Islam yang menekankan tanggung jawab setiap individu untuk memberikan manfaat melalui ilmu dan perbuatan.

Menurut buku "Jawami' al-Kalim: Keindahan Retorika Hadis Nabi Muhammad SAW", disebutkan bahwa di antara anugerah istimewa yang dikaruniakan Allah kepada Nabi Muhammad SAW adalah kemampuan Jawami' al-Kalim. Keistimewaan ini menjadi salah satu bentuk mukjizat yang membedakan beliau dari nabi-nabi lainnya, di mana setiap ungkapan yang beliau sampaikan tidak hanya mengandung makna harfiah, melainkan juga menyimpan hikmah dan makna yang sangat mendalam. Hal ini memungkinkan pesan-pesan beliau dipahami secara luas dan relevan oleh semua kalangan dari berbagai latar belakang (Al-Hajiri, 2018).

Dalam konteks peringkasan teks, proses ini memastikan bahwa hasil ringkasan tetap menyampaikan esensi dan informasi penting dari dokumen asli, tanpa menghilangkan nilai-nilai utamanya. Dengan demikian, konsep Jawami' al-Kalim menjadi inspirasi dalam menciptakan teknologi yang mampu memberikan informasi padat, relevan, dan bermakna.

1.2 Rumusan Masalah

Dengan mempertimbangkan permasalahan dari latar belakang, pokok persoalan dapat disimpulkan seperti ini: "Bagaimana perbandingan kinerja *K-Means Clustering* dan *TextRank* dalam peringkasan teks otomatis berita berbahasa Indonesia menggunakan metrik ROUGE-1, ROUGE-2, dan ROUGE-L untuk mendapatkan nilai *precision, recall, f1-score* pada berbagai tingkat kompresi?"

1.3 Batasan Masalah

Guna menjaga fokus terhadap permasalahan inti, maka cakupan penelitian ini dibatasi pada:

1. Penelitian ini menggunakan dataset *Indonesian Text Summarization* (IndoSum).
2. Penelitian ini berfokus pada peringkasan ekstraktif dengan inputan *single document*.

1.4 Tujuan Penelitian

Penelitian berikut diarahkan untuk hal analisis dalam mengkaji perbandingan kinerja metode *K-Means Clustering* dan *TextRank* dalam peringkasan teks otomatis berita berbahasa Indonesia, serta mengevaluasi efektivitasnya pada berbagai tingkat kompresi menggunakan metrik ROUGE-1, ROUGE-2, dan ROUGE-L (*precision*, *recall*, dan *f1-score*).

1.5 Manfaat Penelitian

1. Melalui penelitian ini diharapkan dapat membantu meningkatkan efektivitas sistem peringkasan dokumen otomatis, terfokus pada berita berbahasa Indonesia, dengan mengidentifikasi kelebihan dan kelemahan masing-masing metode pada berbagai tingkat kompresi. Diharapkan hasil perbandingan ini dapat membantu pengembangan sistem peringkasan yang lebih seimbang antara cakupan informasi dan ketepatan pemilihan kalimat.
2. Hasil analisis perbandingan kinerja kedua metode dan pengaruh tingkat kompresi pada metrik *precision*, *recall*, dan *f1-score* membuka peluang

untuk penelitian lanjutan yang dapat mengeksplorasi parameter lain dalam masing-masing metode, serta membandingkan kedua pendekatan ini dengan metode peringkasan otomatis lainnya untuk berita berbahasa Indonesia.

BAB II

STUDI PUSTAKA

2.1 Penelitian Terkait

Penulis mengangkat topik penelitian dengan judul "Analisis Perbandingan *K-Means Clustering* dan *Textrank* dalam Peringkasan Teks Berita Berbahasa Indonesia" yang merupakan penelitian berkelanjutan dari kajian-kajian sebelumnya di bidang peringkasan teks.

Penelitian oleh Abdillah, (2024) mengkaji perbandingan metode *Textrank* dan *Long Short Term Memory* untuk meringkas teks berita dalam bahasa Inggris. Dari hasil evaluasi terhadap 13.760 data berita, metode *Textrank* menunjukkan kinerja rata-rata dengan nilai *recall* 0.383, *precision* 0.608, dan *f1-score* 0.469. Untuk performa terendah, *Textrank* mencatat *recall* 0.044, *precision* 0.068, dan *f1-score* 0.054, sementara nilai tertinggi ketiga metrik tersebut mencapai sekitar 1.0. Dalam pengaturan pembagian dataset dengan proporsi 90% untuk pelatihan dan 10% untuk pengujian, metode LSTM menunjukkan performa terbaik dengan capaian *recall* 0.462, *precision* 0.507, dan *f1-score* 0.480.

Penelitian oleh Ahsan, (2023) mengeksplorasi peringkasan teks multi-dokumen berita bahasa Indonesia dengan memanfaatkan *FastText* dan *K-Means Clustering*. Penelitian ini menggunakan dataset berisi 30 multi-dokumen dengan 4 kategori label yang beragam. Dari keempat variasi label tersebut, sistem peringkasan mencapai performa optimal saat dibandingkan dengan variasi pertama, namun menunjukkan kinerja terendah ketika dibandingkan dengan variasi keempat.

Variasi pertama menghasilkan nilai rata-rata *precision* 0,674, *recall* 0,620, dan *f-measure* 0,637. Sementara itu, variasi keempat mencatatkan rata-rata *precision* 0,382, *recall* 0,790, dan *f-measure* 0,505.

Penelitian oleh Fadhila dan Nuryana, (2024) mengkaji peringkasan teks otomatis pada artikel portal berita CNN Indonesia dengan menerapkan algoritma *TextRank*. Dari evaluasi terhadap 25 artikel berita berbahasa Indonesia yang berasal dari portal CNN Indonesia, penerapan algoritma *TextRank* berhasil menghasilkan ringkasan dengan rata-rata tingkat kompresi sebesar 6,9% lebih singkat dibandingkan artikel aslinya, dengan pembatasan *output* ringkasan maksimal 3 kalimat per artikel.

Penelitian oleh Prabowo et al. (2017) mengeksplorasi peringkasan teks ekstraktif untuk literatur ilmu komputer berbahasa Indonesia dengan menggunakan pendekatan *Normalized Google Distance* dan *K-Means*. Metode yang diajukan telah diuji validitasnya menggunakan *Rouge score* pada dataset *benchmark*. Hasil penelitian menunjukkan bahwa teknik peringkasan teks yang menggabungkan NGD dan *K-means* meraih nilai rata-rata optimal untuk metrik *precision*, *recall*, dan *relative utility* masing-masing pada evaluator pertama (0,20, 0,47, 0,48) dan evaluator kedua (0,27, 0,43, 0,45). Sementara itu, nilai rata-rata kappa yang diperoleh sebesar 0,41 yang termasuk dalam kategori *moderate*.

Penelitian oleh Abdurrohman, (2018) mengkaji evaluasi algoritma *Textrank* untuk peringkasan teks dalam bahasa Indonesia. Proses penilaian dilakukan dengan menggunakan metrik ROUGE melalui perbandingan antara hasil *Textrank* dengan ringkasan yang disusun secara manual oleh ahli bahasa Indonesia. Algoritma

Textrank mencatatkan nilai rata-rata *f-score* sebesar 0,439 untuk ROUGE-1 dan 0,3186 untuk ROUGE-2.

Penelitian oleh Hernawan et al. (2022) mengeksplorasi peringkasan artikel bahasa Indonesia dengan menerapkan *Textrank* yang diperkuat pembobotan BM25. Melalui perbandingan antara hasil sistem peringkasan otomatis dengan ringkasan yang dibuat oleh pakar pada 10 dokumen, penelitian ini meraih kualitas peringkasan terbaik saat menerapkan *compression rate* 30% dengan capaian nilai rata-rata *precision*, *recall*, dan *f-measure* berturut-turut sebesar 0,552; 0,552; dan 0,552.

Penelitian oleh Samosir et al. (2022) melakukan penelitian tentang peringkasan ekstraktif menggunakan model BERT yang dikombinasikan dengan teknik *K-Means Clustering* untuk merangkum dokumen ilmiah, dengan menggunakan dataset koleksi makalah akademis dari konferensi NeurIPS. Evaluasi performa sistem dilakukan menggunakan dua metrik utama yaitu ROUGE-L yang mencapai skor 15,52% dan BERT *Score* dengan nilai 85,55%, dimana hasil tersebut menunjukkan performa yang lebih unggul dibandingkan dengan metode-metode sebelumnya seperti *PyTextRank* dan BERT *Extractive Summarizer*.

Pada tabel 2.1 akan dipaparkan lebih lanjut mengenai penelitian-penelitian terkait yang dijadikan sumber dan acuan peneliti untuk melakukan penelitian Analisis Perbandingan *K-Means Clustering* dan *Textrank* dalam Peringkasan Teks Berita Berbahasa Indonesia.

Tabel 2.1 Penelitian Terkait

No	Peneliti	Judul	Metode	Hasil
1	(Abdillah, 2024)	Analisis Perbandingan *Textrank* dan *Long Short-Term Memory* Dalam Peringkasan	*TextRank* dan *Long Short-Term Memory*	Pada metode *TextRank* rata-rata nilai *recall*, *precision*, dan *f1-score* yang dihasilkan dari pengujian 13.760 data berita secara berurutan adalah

		Teks Berita Bahasa Inggris		0.383, 0.608, dan 0.469. Metode LSTM menghasilkan nilai rata-rata tertinggi untuk *recall*, *precision*, dan *f1-score* dalam skenario 90% data *training* dan 10% data *testing*, sebesar 0.462, 0.507, dan 0.480
2	(Ahsan, 2023)	Peringkasan Teks Multi Dokumen Berita Berbahasa Indonesia Menggunakan *Fasttext* Dan *K-Means Clustering*	*Fasttext* Dan *K-Means Clustering*	Analisis terhadap 30 data multi-dokumen dengan 4 klasifikasi label menghasilkan temuan bahwa variasi 1 mengunggul dengan rata-rata *precision* 0,674, *recall* 0,620, dan *f-measure* 0,637. Berbeda dengan variasi 4 yang memperoleh rata-rata *precision* 0,382, *recall* 0,790, dan *f-measure* 0,505
3	(Fadhila & Nuryana, 2024)	Teks Ringkas Otomatis pada Portal Berita CNN Indonesia Menggunakan Algoritma *Textrank*	*Textrank*	Ringkasan berita CNN Indonesia menggunakan algoritma *TextRank* mendapatkan hasil rata-rata presentase ringkasan sebesar 6,9% lebih ringkas daripada artikel asli, dengan catatan penulis membatasi jumlah kalimat sebesar 3 kalimat saja.
4	(Prabowo et al., 2017)	Peringkasan Teks Ekstraktif Kepustakaan Ilmu Komputer Bahasa Indonesia Menggunakan Metode *Normalized Google Distance* Dan *K-Means*.	*Normalized Google Distance* dan *K-Means*.	Pendekatan peringkasan teks dengan metode NGD dan *K-means* menunjukkan hasil optimal pada rata-rata akurasi *precision*, *recall*, dan *relative utility* masing-masing 0,20, 0,47, 0,48 menurut ahli pertama dan 0,27, 0,43, 0,45 menurut ahli kedua. Nilai rata-rata *kappa* yang dicapai adalah 0,41 dengan level *moderate*
5	(Abdurrohman, 2018)	Evaluasi Algoritma *Textrank* Pada Peringkasan Teks Berbahasa Indonesia.	*Textrank*	Pengujian menggunakan metrik ROUGE diimplementasikan melalui perbandingan *output* peringkasan *Textrank* dengan *ground truth* ringkasan manual dari pakar bahasa Indonesia. Algoritma *Textrank* mendemonstrasikan rata-rata *f1-score* sebesar 0,439 untuk ROUGE-1 dan 0,3186 untuk ROUGE-2
6	(Hernawan et al., 2022)	Peringkasan Artikel Berbahasa Indonesia	*TextRank* dan BM25	Penelitian ini berhasil dilakukan dengan kualitas

		Menggunakan *TextRank* dengan Pembobotan BM25		ringkasan terbaik didapatkan pada saat penggunaan *compression rate* sebesar 30% dengan nilai rata-rata *precision*, *recall*, dan *f-measure* secara berturut-turut adalah 0,552; 0,552; dan 0,552.
7	(Samosir et al., 2022)	Peringkasan Ekstraktif BERT Dengan *K-Means Clustering* Pada Karya Ilmiah	*K-Means Clustering* dan BERT	Evaluasi kinerja dilakukan dengan ROUGE-L memberikan hasil sebesar 15,52% dan *BERTScore* sebesar 85,55%.

Penelitian ini memiliki perbedaan signifikan dibandingkan penelitian sebelumnya baik dari segi dataset, metode, maupun hasil yang diharapkan. Penelitian ini memanfaatkan dataset IndoSum sebagai sumber data, yang berisi teks-teks dalam bahasa Indonesia dengan penekanan pada pembuatan ringkasan yang representatif. Hal ini berbeda dengan penelitian oleh (Fadhila & Nuryana, 2024), yang menggunakan algoritma *TextRank* untuk merangkum berita dari portal CNN Indonesia, serta penelitian oleh (Ahsan, 2023) yang menggabungkan metode *FastText* dan *K-Means Clustering* untuk menghasilkan ringkasan teks multi-dokumen berita Indonesia.

Selain itu, penelitian ini membandingkan dua metode berbeda, yaitu *K-Means Clustering* dan *TextRank*, untuk menganalisis metode mana yang menghasilkan ringkasan teks yang lebih optimal. Evaluasi kedua metode ini memberikan analisis yang lebih komprehensif dibandingkan penelitian sebelumnya. Penelitian oleh (Prabowo et al., 2017), memanfaatkan *Normalized Google Distance* dan *K-Means* untuk pendekatan ekstraktif tanpa mengintegrasikan algoritma berbasis grafik seperti *TextRank*. Penelitian oleh (Abdurrohman, 2018)

juga berbeda karena hanya menggunakan *TextRank* untuk evaluasi peringkasan teks berdasarkan metrik ROUGE.

Kelebihan dari pendekatan penelitian ini juga dipengaruhi inspirasi penelitian (Hernawan et al., 2022) , yang menggunakan kombinasi *TextRank* dan BM25 untuk menghasilkan ringkasan artikel dengan skor ROUGE yang tinggi pada dataset tertentu. Penelitian ini melakukan perbandingan antara *pendekatan K-Means Clustering* dengan *Textrank* untuk menemukan metode yang lebih efektif dalam menciptakan ringkasan teks yang akurat dan representatif, terutama ketika diterapkan pada dataset IndoSum. Berdasarkan penelitian terkait, dapat diamati bahwa berbagai pendekatan telah digunakan dalam peringkasan teks otomatis, masing-masing dengan keunggulan dan keterbatasan tersendiri. Analisis perbandingan spesifik antara *K-Means Clustering* dan *TextRank* untuk peringkasan teks berita berbahasa Indonesia masih relatif jarang dieksplorasi dalam konteks dataset IndoSum. Mempertimbangkan karakteristik *K-Means Clustering* dalam mengelompokkan kalimat berdasarkan kesamaan semantik dan fitur *TextRank* dalam mengidentifikasi kalimat penting berdasarkan struktur graf, peneliti termotivasi untuk membandingkan kinerja kedua metode tersebut untuk peringkasan teks otomatis berita berbahasa Indonesia. Analisis perbandingan ini diharapkan dapat mengidentifikasi metode mana yang lebih unggul dalam menghasilkan ringkasan yang komprehensif pada berbagai kondisi dan tingkat kompresi.

2.2 Peringkasan Teks Otomatis

Teknik peringkasan teks otomatis atau yang dikenal dengan *Automatic Text Summarization* merupakan metode untuk mengekstrak informasi utama dari suatu dokumen atau teks dengan memanfaatkan teknologi komputer (Halimah et al., 2022). Adapun tujuan dari peringkas teks ini adalah untuk membantu pembaca dalam menemukan inti utama informasi dari teks yang dibaca tanpa membaca keseluruhan teksnya (Andriani et al., 2019). Metode peringkasan teks dapat diklasifikasikan ke dalam dua kategori utama, yaitu pendekatan ekstraktif dan abstraktif. Metode ekstraktif bekerja dengan cara memilih dan mengambil kalimat-kalimat yang dianggap penting dari dokumen sumbernya. Sementara itu, metode abstraktif beroperasi dengan cara mereduksi konten dokumen berdasarkan makna semantik, sehingga memerlukan implementasi *Natural Language Processing* dalam prosesnya (Nagalavi & Hanumanthappa, 2019).

Peringkasan teks otomatis merupakan bagian dari pemrosesan bahasa alami *Natural Language Processing* yang berfungsi untuk meringkas teks melalui pengurangan panjang dokumen sambil tetap menjaga inti informasi yang terkandung di dalamnya. Metode ini sangat berguna dalam era digital saat ini, dimana jumlah informasi yang tersedia sangat besar dan sering kali membutuhkan waktu yang signifikan untuk dibaca secara menyeluruh. Teknologi ini bisa digunakan dengan berbagai algoritma, mulai dari metode statistik sederhana hingga model *machine learning* yang kompleks.

Jika dilihat lebih lanjut berdasarkan inputannya, sistem peringkasan otomatis ini dilakukan dengan menggunakan *single-document* dimana teks yang

diringkas berasal dari satu dokumen saja untuk menghasilkan ringkasan menyeluruh. Dengan menggunakan teknik ini, esensi dari teks dapat diperoleh secara cepat tanpa harus membaca seluruh konten asli, sehingga memudahkan pengguna dalam memahami informasi utama yang disampaikan.

2.3 *Natural Language Processing*

Teknik *Natural Language Processing* (NLP) merupakan bagian dari bidang ilmu komputer yang merupakan percabangan dari *Artificial Intelligence* dan bahasa (linguistik) serta memiliki keterkaitan dengan interaksi antara komputer dan bahasa alami manusia. Sasaran utama dari NLP adalah mengembangkan sistem mesin yang dapat memahami dan menginterpretasi makna bahasa serta mampu memberikan tanggapan yang tepat (Alamanda et al., 2016). Selain itu, NLP juga memungkinkan komputer untuk melakukan tugas-tugas seperti pengenalan suara, klasifikasi teks bahkan ekstraksi informasi dengan tingkat akurasi yang cukup tinggi.

Cara kerja *Natural Language Processing* (NLP) melibatkan beberapa tahapan penting untuk memproses dan memahami teks dalam bahasa alami. Beberapa tahapan yang dilakukan antara lain *segmentation* untuk memisahkan teks asli setiap artikel pada dataset berdasarkan paragraf. Setelah itu, terdapat tahapan *preprocessing* seperti *case folding, noise removal, stopword removal* dan *stemming* yang dilakukan untuk mempersiapkan dataset sebelum diolah menggunakan model algoritma yang digunakan. Setelah tahap *preprocessing* selesai, Teknik pembobotan kata seperti T*erm Frequency-Inverse Document Frequency* (TF-IDF) diterapkan untuk mengukur tingkat kepentingan kata-kata di dalam suatu dokumen teks. Tahap akhir melibatkan penggunaan algoritma *machine learning* atau *deep*

learning untuk membuat hasil peringkasan atau mengekstrak informasi dari teks tersebut. *Natural Language Processing* juga berperan penting dalam pengembangan teknologi seperti *chatbots*, penerjemah otomatis, analisis teks, dan peringkasan teks yang dapat digunakan dalam berbagai kegiatan berskala besar.

2.4 *K-Means Clustering*

Clustering atau pengelompokkan merupakan salah satu teknik fundamental dalam data mining yang bertujuan untuk mengorganisir data ke dalam kelompok-kelompok berdasarkan kesamaan karakteristik tertentu. Dalam konteks pengembangan sistem peringkasan teks otomatis, teknik clustering menjadi sangat penting untuk mengelompokkan kalimat-kalimat yang memiliki kemiripan makna atau topik, sehingga dapat menghasilkan ringkasan yang lebih terstruktur dan representatif. *K-Means Clustering* merupakan salah satu algoritma *clustering* yang paling populer dan banyak digunakan karena kesederhanaannya dalam implementasi serta efektivitasnya dalam menangani dataset berukuran besar.

Penerapan algoritma *K-Means* dalam sistem peringkasan teks tidak hanya dilihat dari aspek teknis semata, tetapi juga mencerminkan nilai-nilai Islam yang mengajarkan tentang keteraturan dan hikmah dalam mengorganisir sesuatu. Konsep pengelompokkan yang sistematis ini sejalan dengan ajaran Islam yang mengakui keberagaman dan klasifikasi sebagai bagian dari sunnatullah dalam penciptaan alam semesta. Oleh karena itu, pembahasan mengenai *K-Means Clustering* dalam penelitian ini akan diuraikan melalui dua perspektif, yaitu aspek teoritis algoritma dan kaitannya dengan nilai-nilai Islam dalam pengelompokkan.

2.4.1 Teori *K-Means Clustering*

Algoritma *K-Means Clustering* ialah suatu metode yang berfungsi untuk mengelompokkan data menjadi beberapa *cluster* berdasarkan kesamaan karakteristik atau atribut yang dimiliki (Snyder, 2019). Algoritma ini diusulkan oleh Stuart Lloyd pada tahun 1957 dan diperkenalkan kembali oleh James MacQueen pada tahun 1967 sembari diberi nama *K-Means Clustering.* Sasaran dari algoritma tersebut adalah meminimalisir fungsi objektif yang diterapkan dalam proses *clustering*. Pada dasarnya, metode ini berusaha untuk memperkecil variabilitas di dalam setiap cluster sambil memperbesar variabilitas di antara *cluster* yang berbeda (Abdussalam Amrullah et al., 2022).

Metode kerja algoritma ini dimulai dengan menetapkan jumlah *cluster* atau nilai k pada tahap awal. Selanjutnya, dilakukan perhitungan jarak antara kata dengan *centroid*. Untuk mengukur jarak tersebut, digunakan rumus dengan pendekatan *Euclidean distance* sebagai berikut:

$$d(x_i, \mu_k) = \sqrt{\sum_{j=1}^{n} \left(x_{ij} - \mu_{kj}\right)^2} \qquad (2.1)$$

Keterangan:
$d(x_i, \mu_k)$ = jarak *euclidean* antara vektor TF-IDF kalimat x_i dan *centroid* μ_k dari kluster ke-k
x_i = vektor TF-IDF kalimat ke-i dalam satu dokumen
μ_k = *centroid* (pusat) dari kluster ke-k
j = indeks fitur dalam vektor vektor TF-IDF
n = jumlah dimensi fitur (jumlah kata unik atau fitur dalam vektor TF-IDF)
x_{ij} = nilai vektor TF-IDF untuk kalimat i pada dimensi fitur ke- j
μ_{kj} = nilai *centroid* kluster *k* pada dimensi fitur ke- j

Euclidean Distance merupakan sebuah metode pengukuran jarak antara dua titik dalam ruang vektor berdasarkan panjang garis lurus yang menghubungkan

keduanya. Dalam konteks pengelompokan seperti *K-Means Clustering*, *euclidean distance* digunakan untuk menghitung jarak antara vektor kalimat (x_i) dan *centroid cluster* (μ_i) guna menentukan kedekatan kalimat terhadap suatu *cluster*. Rumusnya dinyatakan sesuai pada Persamaan 2.1. Nilai jarak ini digunakan untuk menentukan keanggotaan kalimat dalam *cluster* tertentu berdasarkan *centroid* dengan jarak terdekat.

Setelah memperoleh hasil kalkulasi jarak, maka dapat diketahui bahwa suatu data akan diklasifikasikan ke dalam *cluster* tertentu berdasarkan jarak terpendek dari seluruh *cluster* yang ada. Nilai *centroid* awal akan dibentuk secara random pada setiap iterasi, yang kemudian akan diperbarui dengan nilai terbaru dari *centroid* menggunakan Persamaan 2.2 berikut ini:

$$\mu_k = \frac{1}{|\boldsymbol{C_k}|} \sum_{x_i \in \boldsymbol{C_k}} x_i \qquad (2.2)$$

Keterangan:

μ_k	= *centroid* (pusat) dari *cluster* k
C_k	= himpunan kalimat (vektor TF-IDF x_i) yang termasuk dalam kluster k
$\|\boldsymbol{C_k}\|$	= jumlah kalimat dalam kluster k
x_i	= vektor TF-IDF kalimat i yang berada dalam kluster k
$\sum_{x_i \in \boldsymbol{C_k}} x_i$	= penjumlahan semua vektor TF-IDF x_i dalam kluster C_k

K-Means Clustering melakukan pembaharuan *centroid* dengan menghitung rata-rata posisi semua data yang termasuk dalam suatu *cluster*. Proses ini bertujuan untuk memposisikan *centroid* secara lebih representatif terhadap anggotanya pada iterasi berikutnya. Rumus pembaharuan *centroid* dituliskan pada Persamaan 2.2 dengan menghitung rata-rata semua vektor data dalam *cluster*, posisi *centroid* akan berpindah ke lokasi yang lebih optimal, sehingga menghasilkan partisi *cluster* yang lebih akurat seiring iterasi algoritma. Proses pengulangan akan terus berjalan

sampai posisi *centroid* tidak mengalami perubahan yang berarti atau sudah mencapai batas maksimum iterasi yang telah ditetapkan.

2.4.1 Perspektif Islam Tentang Pengelompokan

Konsep pengelompokkan dalam teknologi *clustering* mencerminkan nilai-nilai Islam yang mengajarkan tentang keteraturan dan hikmah dalam penciptaan Allah SWT. Sebagai muslim yang mengembangkan teknologi, implementasi algoritma *K-Means* tidak hanya bertujuan untuk efisiensi teknis, tetapi juga sebagai bentuk ketaatan dalam menerapkan ilmu pengetahuan yang bermanfaat bagi umat manusia. Ketaatan tersebut dilandaskan pada kewajiban kita sebagai manusia untuk senantiasa menaati perintah Allah SWT dalam mengembangkan teknologi yang tidak memecah belah, melainkan menyatukan dan mengorganisir data secara harmonis. Pendekatan ini selaras dengan firman Allah SWT dalam Al-Qur'an Surah Al-An'am ayat 159:

اِنَّ الَّذِيْنَ فَرَّقُوْا دِيْنَهُمْ وَكَانُوْا شِيَعًا لَّسْتَ مِنْهُمْ فِيْ شَيْءٍۗ اِنَّمَآ اَمْرُهُمْ اِلَى اللّٰهِ ثُمَّ يُنَبِّئُهُمْ بِمَا كَانُوْا يَفْعَلُوْنَ

"*Sesungguhnya orang-orang yang memecah belah agamanya dan mereka menjadi (terpecah) dalam golongan-golongan, sedikit pun engkau (Nabi Muhammad) tidak bertanggung jawab terhadap mereka. Sesungguhnya urusan mereka (terserah) hanya kepada Allah. Kemudian, Dia akan memberitahukan kepada mereka apa yang telah mereka perbuat.*" *(QS.Al-An'am:159)*

Ayat tersebut menjelaskan tentang pertanggungjawaban kepada Allah bagi mereka yang memecah belah agama dalam golongan-golongan. Ayat dari Al-Qur'an tersebut berbicara tentang kelompok-kelompok yang terpecah dalam agama, menggambarkan pengelompokan manusia berdasarkan perbedaan pemahaman, sejalan dengan konsep dasar algoritma *K-Means Clustering* yang digunakan dalam

penelitian yang bekerja dengan mengelompokkan teks berita ke dalam kelompok-kelompok berbeda berdasarkan karakteristik tertentu, sama seperti ayat yang menggambarkan manusia yang terkelompok dalam golongan-golongan yang berbeda (Lajnah Pentashihan Mushaf Al-Qur'an, 2019).

Dalam Tafsir Al-Muyassar dijelaskan mengenai orang-orang yang terpecah belah menjadi golongan-golongan, sesuai dengan penelitian ini yang menggunakan *K-Means Clustering*, terdapat keselarasan yang mendalam dengan konsep pengelompokan dalam pandangan Islam. Sebagaimana dijelaskan dalam Tafsir Al-Muyassar, ayat ini membedakan antara pengelompokan berdasarkan golongan yang berbuat kebaikan dan keburukan. Metode *K-Means Clustering* menerapkan konsep pengelompokan yang sejalan dengan ajaran Islam, dimana kalimat-kalimat dikelompokkan berdasarkan kriteria objektif berupa kesamaan semantik. Algoritma ini memetakan kalimat-kalimat ke dalam *cluster* berdasarkan kedekatan karakteristiknya, kemudian memilih yang paling representatif dari setiap kelompok, mencerminkan bagaimana Islam mengajarkan untuk mengambil dan memilih yang terbaik dari setiap kelompok golongan (UIN Sayyid Ali Rahmatullah Tulungagung, 2023).

2.5 *Textrank*

Algoritma *Textrank* adalah pengembangan dari *Pagerank* yang menerapkan pendekatan berbasis graf untuk mengidentifikasi kalimat-kalimat kunci dalam suatu dokumen (Mutlu et al., 2020). *TextRank* menggunakan jaringan dengan sebuah sisi untuk memperlihatkan kesamaan antar kalimat dan jaringan simpul untuk merepresentasikan kalimat pada dokumen. Penggunaan algoritma *TextRank*

digunakan untuk memberikan penarikan keputusan serta mengekstrak kalimat pada sistem peringkas otomatis. Adapun tujuannya adalah untuk mendapatkan semua poin kalimat penting yang ada pada suatu dokumen tersebut.

Mekanisme kerja algoritma *Textrank* dimulai dengan mengidentifikasi dokumen yang akan diproses, kemudian setiap kalimat dimodelkan sebagai *node* dan fungsi *similarity* yang menggambarkan relasi antar kalimat diwakili sebagai *edge*. *Edge* dalam graf dapat bersifat berarah atau tidak berarah. Setelah graf terbentuk, dilakukan proses pemeringkatan graf, dan selanjutnya *node* dengan peringkat tertinggi akan dipilih berdasarkan skor peringkat kalimat tersebut (Hernawan et al., 2022).

Untuk membangun graf, langkah awal adalah menghitung bobot sisi antar simpul yang merepresentasikan hubungan antar kalimat. Bobot ini dihitung berdasarkan tingkat kemiripan (*similarity*) antar kalimat dalam dokumen. Salah satu pendekatan umum untuk menghitung kemiripan ini adalah menggunakan *cosine similarity*, yang mengukur hubungan antar vektor kalimat dengan menghitung sudut *cosinus* antara dua vektor TF-IDF. *Cosine similarity* digunakan karena mampu menangkap hubungan linear antara elemen-elemen dalam vektor sehingga memberikan hasil yang lebih representatif. Pada Persamaan 2.3 akan dipaparkan rumus untuk *Cosine similarity*

$$Sim(x_i, x_j) = \frac{\sum_{k=1}^{n} (x_{ik} \cdot x_{jk})}{\sqrt{\sum_{k=1}^{n} (x_{ik})^2} \cdot \sqrt{\sum_{k=1}^{n} (x_{jk})^2}} \quad (2.3)$$

Keterangan:

$Sim(x_i, x_j)$ = nilai kemiripan kosinus antara vektor TF-IDF kalimat x_i dan x_j

x_i, x_j = vektor TF-IDF dari kalimat i dan kalimat j dalam dokumen

n = jumlah dimensi fitur (jumlah kata unik atau fitur dalam vektor TF-IDF)

x_{ik} = nilai vektor TF-IDF dari kalimat i pada fitur ke-k
x_{jk} = nilai vektor TF-IDF dari kalimat j pada fitur ke-k
$\sum_{k=1}^{n}(x_{ik} \cdot x_{jk})$ = penjumlahan hasil perkalian nilai vektor TF-IDF pada setiap dimensi fitur antara x_i dan x_j
$\sqrt{\sum_{k=1}^{n}(x_{ik})^2}$ = panjang (*magnitude*) dari vektor TF-IDF x_i
$\sqrt{\sum_{k=1}^{n}(x_{jk})^2}$ = panjang (*magnitude*) dari vektor TF-IDF x_j

Cosine similarity menghasilkan skor dalam rentang 0 hingga 1, dengan nilai yang lebih tinggi mengindikasikan kesamaan yang lebih besar di antara kedua kalimat tersebut. Dengan bobot sisi yang dihitung menggunakan *cosine similarity*, graf *TextRank* menjadi lebih akurat dalam merepresentasikan hubungan antar kalimat. Hasil perhitungan ini menjadi dasar untuk langkah selanjutnya dalam algoritma, yaitu penghitungan peringkat simpul menggunakan *PageRank*. Setelah graf dengan bobot *cosine similarity* terbentuk, algoritma *PageRank* digunakan untuk menghitung peringkat setiap simpul (kalimat) dalam graf. Adapun Persamaan 2.4 merupakan rumus *PageRank*

$$S(V_i) = (1-d) + d \cdot \sum_{V_j \in \mathrm{In}(V_i)} \frac{S(V_j)}{Out(V_j)} \quad (2.4)$$

Keterangan:
$S(V_i)$ = skor penting dari simpul (kalimat) V_i dalam graf
d = faktor peredaman (*damping factor*)
V_j = simpul lain yang memiliki hubungan sisi ke V_i
$In(v_i)$ = himpunan simpul yang memiliki sisi menuju V_i
$Out(V_j)$ = jumlah sisi yang keluar dari simpul V_j
$S(V_j)$ = skor *PageRank* dari simpul V_j yang terhubung dengan V_i

Setelah skor *PageRank* dihitung untuk setiap simpul dalam graf, nilai-nilai ini digunakan untuk menentukan urutan prioritas kalimat dalam dokumen. Simpul dengan skor tertinggi dianggap sebagai kalimat paling penting, karena memiliki hubungan yang kuat dengan banyak simpul lainnya atau dengan simpul yang

memiliki skor tinggi. Dalam konteks peringkasan teks, kalimat-kalimat ini dipilih secara bertahap sesuai dengan skor untuk membentuk ringkasan akhir. Mekanisme ini menjamin bahwa kalimat yang dipilih tidak hanya bermakna pada tingkat lokal, melainkan juga memberikan dampak menyeluruh terhadap dokumen, berdasarkan arsitektur graf yang telah dibentuk. Dengan demikian, algoritma *TextRank* memberikan ringkasan yang representatif terhadap isi dokumen, mengutamakan kalimat-kalimat yang paling relevan dan signifikan.

2.6 Peringkasan Teks Ekstraktif

Peringkasan teks ekstraktif merupakan metode yang mengambil kalimat atau frasa langsung dari teks asli tanpa melakukan modifikasi atau pembuatan ulang kalimat. Metode ini berfokus pada pengambilan komponen-komponen krusial dari teks sambil tetap mempertahankan kosakata dan format kalimat yang ada dalam dokumen sumber. Pada proses ini, setiap kalimat dalam dokumen dianalisis berdasarkan relevansinya terhadap isi keseluruhan teks. Hasil ringkasan berupa gabungan kalimat-kalimat yang paling representatif dari dokumen asli. Teknik ini banyak digunakan karena lebih sederhana dan tidak memerlukan pemahaman mendalam terhadap konteks semantik seperti halnya pada peringkasan abstraktif.

Pendekatan ekstraktif umumnya memanfaatkan teknik pemrosesan bahasa alami (NLP) untuk mengkaji dan mengevaluasi tingkat kepentingan setiap kalimat dalam teks. Salah satu metode yang diterapkan yaitu penimbangan kalimat menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF), yang menghitung tingkat kemunculan kata dalam dokumen tertentu dan

membandingkannya dengan frekuensi kata tersebut pada keseluruhan koleksi dokumen.

Lebih lanjut, pendekatan lain seperti *K-Means Clustering* dan *Textrank* dapat dimanfaatkan untuk mengidentifikasi kalimat-kalimat penting berdasarkan relasi antar kalimat dalam dokumen. *K-Means Clustering* mengelompokkan kalimat berdasarkan kesamaan vektornya. Dalam hal ini, kalimat yang paling dekat dengan *centroid* suatu *cluster* dipilih sebagai representasi dari *cluster* tersebut. Sementara itu, *TextRank* membangun grafik dengan simpul-simpul berupa kalimat dan bobot antar simpul dihitung menggunakan metrik seperti *cosine similarity*. Dengan algoritma *PageRank*, setiap simpul diberikan skor relevansi berdasarkan bobot hubungannya dengan simpul lainnya, dan kalimat-kalimat dengan skor tertinggi dipilih sebagai ringkasan.

Dalam penelitian ini, metode ekstraktif digunakan untuk menghasilkan ringkasan dari teks berita berbahasa Indonesia dengan memilih kalimat-kalimat penting berdasarkan analisis bobot kata. Kalimat-kalimat tersebut dipilih tanpa mengubah kata atau struktur kalimat, sehingga ringkasan yang dihasilkan terdiri dari elemen-elemen yang ada dalam teks asli.

Metode ekstraktif memiliki beberapa kelebihan, seperti kemudahan dalam implementasi dan kemampuan untuk menghasilkan ringkasan yang tetap akurat karena tidak ada modifikasi pada isi kalimat. Namun, kelemahannya adalah metode ini cenderung menghasilkan ringkasan yang kurang fleksibel dibandingkan dengan metode abstraktif, karena tidak dapat melakukan penggabungan atau penyusunan ulang informasi untuk menghasilkan ringkasan yang lebih halus dan koheren.

2.7 *Recall Oriented Understudy for Gisting Evaluation*

ROUGE *Recall-Oriented Understudy for Gisting Evaluation* merupakan sistem pengukuran yang berfungsi untuk mengukur mutu peringkasan teks melalui perbandingan antara ringkasan yang dihasilkan sistem otomatis dengan ringkasan acuan yang dibuat manusia. ROUGE menjadi salah satu metrik evaluasi yang paling banyak diterapkan dalam bidang peringkasan teks, karena dapat memberikan indikasi sejauh mana ringkasan otomatis mendekati versi referensi berdasarkan kesamaan leksikal, struktur frasa, dan susunan kalimat.

ROUGE terdiri dari beberapa varian, yang paling sering digunakan adalah ROUGE-N. ROUGE-N mengukur kesamaan antara n-gram (*unigram, bigram*, dan seterusnya) yang ada dalam ringkasan otomatis dan referensi. Seperti contoh, ROUGE-1 mengukur seberapa banyak kata tunggal (*unigram*) dalam ringkasan otomatis yang sesuai dengan ringkasan manual, sementara ROUGE-2 mengukur kesamaan dalam bentuk pasangan kata berurutan (*bigram*). *Precision*, *Recall*, dan *F-Measure* dihitung untuk menentukan rasio kata-kata yang benar-benar berkaitan dengan ringkasan acuan.

Varian berikutnya yaitu ROUGE-L, yang berfungsi untuk menghitung *Longest Common Subsequence* (LCS) atau rangkaian kata terpanjang yang identik antara ringkasan yang dihasilkan sistem dengan ringkasan referensi. ROUGE-L bermanfaat untuk mengevaluasi ketepatan urutan kata dan keterkaitan keseluruhan kalimat. Metrik ini memungkinkan pemahaman mengenai seberapa efektif urutan informasi dapat dipertahankan dalam ringkasan otomatis. Proses evaluasi dengan

ROUGE umumnya mencakup tiga elemen pokok, yaitu *Precision*, *Recall*, dan *F-Measure*:

1. *Precision* (Presisi) untuk ROUGE : mengukur seberapa banyak n-gram yang dihasilkan oleh sistem cocok dengan n-gram dalam ringkasan referensi.

$$Precision = \frac{Jumlah\ n-gram\ yang\ cocok}{Jumlah\ total\ n-gram\ dalam\ ringkasan\ otomatis} \quad (2.5)$$

2. *Recall* (Relevansi) untuk ROUGE : mengukur seberapa banyak n-gram dari ringkasan referensi yang ditangkap oleh sistem.

$$Recall = \frac{Jumlah\ n-gram\ yang\ cocok}{Jumlah\ total\ n-gram\ dalam\ ringkasan\ manual} \quad (2.6)$$

3. *F-Measure* (*F1-Score*) untuk ROUGE: memberikan keseimbangan antara *Precision* dan *Recall* serta memberikan gambaran umum tentang kualitas ringkasan.

$$F - Measure = 2 \text{ x } \frac{Precision\ x\ Recall}{Precision + Recall} \quad (2.7)$$

Dalam penelitian ini, ROUGE digunakan untuk mengevaluasi kinerja model dalam menghasilkan ringkasan otomatis dari teks ilmiah. Melalui penerapan metrik ROUGE, para peneliti mampu mengevaluasi tingkat akurasi dan kelengkapan ringkasan yang dihasilkan sistem jika dibandingkan dengan ringkasan buatan manusia yang dijadikan sebagai standar acuan. Hasil dari ROUGE akan memberikan nilai evaluasi berdasarkan ROUGE-1, ROUGE-2, dan ROUGE-L untuk mengukur kesamaan kata, frasa, dan urutan kata antara kedua jenis ringkasan

tersebut. Pada tahapan evaluasi menggunakan metrik ROUGE akan memiliki bagian sebagai berikut:

1. ROUGE-N: Menghitung kemiripan n-gram antara ringkasan yang diproduksi sistem dengan ringkasan acuan. Penelitian ini menitikberatkan pada ROUGE-1 (unigram *overlap*) dan ROUGE-2 (bigram *overlap*). ROUGE-1 mengevaluasi jumlah kata tunggal yang serupa antara ringkasan sistem dan referensi, sedangkan ROUGE-2 menganalisis kombinasi dua kata yang berurutan.

$$ROUGE - N = \frac{\text{Jumlah n−gram yang cocok}}{\text{Jumlah total n−gram dalam ringkasan referensi}} \quad (2.9)$$

ROUGE-L: Mengukur *Longest Common Subsequence* (LCS), yang mempertimbangkan kesamaan urutan kata terpanjang yang sama antara ringkasan sistem dan ringkasan referensi. Ini sangat penting untuk menangkap urutan kalimat atau kata yang sama secara keseluruhan.

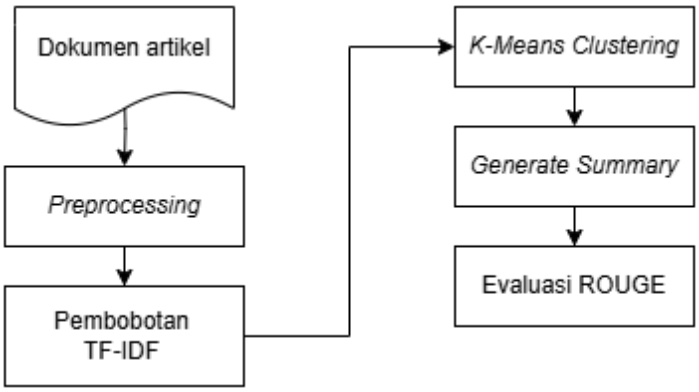
$$ROUGE - L = \frac{\text{Panjang subsequence yang sama terpanjang}}{\text{Panjang total subsequence dalam ringkasan referensi}} \quad (2.10)$$

BAB III

DESAIN DAN IMPLEMENTASI

3.1 Desain Sistem

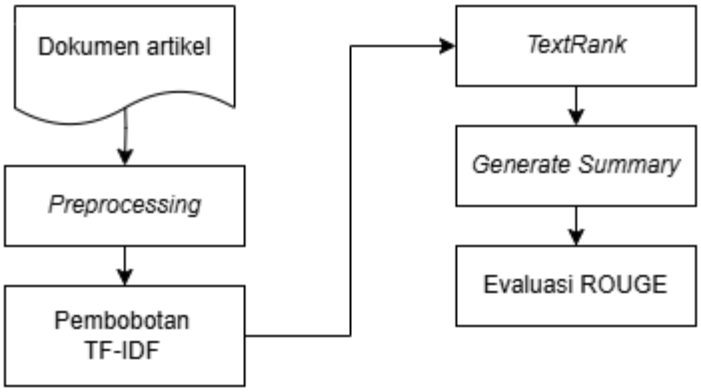
Penelitian ini memiliki dua desain sistem berbeda untuk peringkasan teks otomatis berita berbahasa Indonesia. Kedua desain sistem ini mengimplementasikan metode peringkasan ekstraktif yang berbeda namun dengan tahapan awal yang serupa. Masing-masing desain sistem direpresentasikan dalam Gambar 3.1 dan Gambar 3.2 berikut



Gambar 3.1 Desain Sistem *K-Means Clustering*

Pada Gambar 3.1 terlihat desain sistem peringkasan teks dengan menggunakan pendekatan *K-Means Clustering*. Dalam sistem ini, dokumen artikel akan melalui tahapan *Preprocessing* untuk membersihkan teks, dilanjutkan dengan pembobotan TF-IDF untuk merepresentasikan kalimat dalam bentuk vektor. Hasil pembobotan ini kemudian menjadi *input* untuk proses *K-Means Clustering* yang mengelompokkan kalimat serupa, dilanjutkan dengan *Generate Summary* untuk

membentuk ringkasan, dan diakhiri dengan Evaluasi ROUGE untuk mengukur kualitas ringkasan.



Gambar 3.2 Desain Sistem *TextRank*

Gambar 3.2 menampilkan desain sistem peringkasan teks menggunakan metode *TextRank*. Perbedaan utama dengan desain sistem sebelumnya terletak pada penggunaan algoritma *TextRank* sebagai pengganti *K-Means Clustering*, sementara tahapan awal dan akhirnya tetap sama. *TextRank* memanfaatkan representasi graf untuk menentukan peringkat kalimat berdasarkan hubungan semantiknya dengan kalimat lain dalam dokumen.

3.2 Data Penelitian

Studi ini memanfaatkan dataset yang berasal dari "*Indonesian Text Summarization*" (Indosum) dengan versi terkini yang diperbarui pada tahun 2023. Indosum merupakan sebuah data besar yang terdiri dari kumpulan artikel berita online untuk peringkasan teks berbahasa Indonesia. Dataset Indosum disediakan oleh perusahaan aggregator ringkasan bahasa Indonesia dan berita yang bernama "Shortir". Dataset tersebut terdiri dari kurang lebih 20 ribu artikel berita yang

mencakup judul, kategori, sumber, URL artikel sumber, serta ringkasan yang disusun secara manual oleh dua penutur asli bahasa Indonesia (Kurniawan & Louvan, 2018).

Dataset IndoSum terbagi menjadi tiga bagian pokok, yakni *training*, *development*, dan *testing*, dengan menggunakan format *file.jsonl*. Salah satu *file* dalam dataset, *train.03.jsonl*, berisi 14.290 artikel berita dengan 7 fitur. Salah satu fitur tersebut adalah *paragraphs*, yang berisi paragraf dalam bentuk daftar kalimat dimana setiap kalimat telah dipecah menjadi daftar kata (*token*). Selain itu, fitur *summary* menyediakan ringkasan manual dengan rata-rata 3,48 kalimat untuk setiap artikel. Contoh struktur dari dataset IndoSum yang digunakan dapat dilihat pada Tabel 3.1

Tabel 3.1 Contoh Dataset IndoSum

Category	***Gold_labels***	***Id***	***Paragraphs***	***Source***	***Source_url***	***Summary***
tajuk utama	[[False, True], [True, True], [False, False, False], [False, False], [False, False], [False,	1501893029-lula-kamal-dokter-ryan-thamrin-sakit-sejak-setahun	[[['Jakarta', ',', 'CNN', 'Indonesia', '-', '-', 'Dokter', 'Ryan', 'Thamrin', ',', 'yang', 'terkenal', 'lewat', 'acara', 'Dokter', 'Oz', 'Indonesia', ',', 'meninggal', 'dunia', 'pada', 'Jumat', '(', '4', '/', '8', ')', 'dini', 'hari', '.'], ['Dokter', 'Lula', 'Kamal', 'yang', 'merupakan', 'selebriti', 'sekaligus', 'rekan', 'kerja', 'Ryan', 'menyebut', 'kawannya', 'itu', 'sudah', 'sakit', 'sejak', 'setahun', 'yang', 'lalu', '.']], [['Lula', 'menuturkan', ',', 'sakit', 'itu',	cnn indonesia	https://www.cnnindonesia.com/hiburan/20170804120703-234-232443/lula-kamal-dokter-ryan-thamrin-sakit-sejak-setahun-lalu/	[['Dokter', 'Lula', 'Kamal', 'yang', 'merupakan', 'selebriti', 'sekaligus', 'rekan', 'kerja', 'Ryan', 'Thamrin', 'menyebut', 'kawannya', 'itu', 'sudah', 'sakit', 'sejak', 'setahun', 'yang', 'lalu', '.'], ['Lula', 'menuturkan', ',',

	False], [False, False], [False], [False, False]]		'membuat', 'Ryan', 'mesti', 'vakum', 'dari', 'semua', 'kegiatannya', ',', 'termasuk', 'menjadi', 'pembawa', 'acara', 'Dokter', 'Oz', 'Indonesia', '.'], ['Kondisi', 'itu', 'membuat', 'Ryan', 'harus', 'kembali', 'ke', 'kampung', 'halamannya', 'di', 'Pekanbaru', ',', 'Riau', 'untuk', 'menjalani', 'istirahat', '.']], [['', 'Setahu', 'saya', 'dia', 'orangnya', 'sehat', ',', 'tapi', 'tahun', 'lalu', 'saya', 'dengar', 'dia', 'sakit', '.'], ['(', 'Karena', ')', 'sakitnya', ',', 'ia', 'langsung', 'pulang', 'ke', 'Pekanbaru', ',', 'jadi', 'kami', 'yang', 'mau', 'jenguk', 'juga', 'susah', '.'], ['Barangkali', 'mau', 'istirahat', ',', 'ya', 'betul', 'juga', ',', 'kalau', 'di', 'Jakarta', 'susah', 'isirahatnya', ',', '', 'kata', 'Lula', 'kepada', 'CNNIndonesia.com', ',', 'Jumat', '(', '4', '/', '8', ')', '.']], [['Lula', 'yang', 'mengenal', 'Ryan', 'sejak', 'sebelum', 'aktif', 'berkarier', 'di', 'televisi', 'mengaku', 'belum', 'sempat', 'membesuk', 'Ryan', 'lantaran', 'lokasi', 'yang', 'jauh', '.'], ['Dia', 'juga', 'tak', 'tahu', 'penyakit', 'apa', 'yang', 'diderita', 'Ryan', '.']], [['', 'Itu', 'saya',			'sakit', 'itu', 'membuat', 'Ryan', 'mesti', 'vakum', 'dari', 'semua', 'kegiatannya', ',', 'termasuk', 'menjadi', 'pembawa', 'acara', 'Dokter', 'Oz', 'Indonesia', '.'], ['Kondisi', 'itu', 'membuat', 'Ryan', 'harus', 'kembali', 'ke', 'kampung', 'halamannya', 'di', 'Pekanbaru', ',', 'Riau', 'untuk', 'menjalani', 'istirahat', '.']]

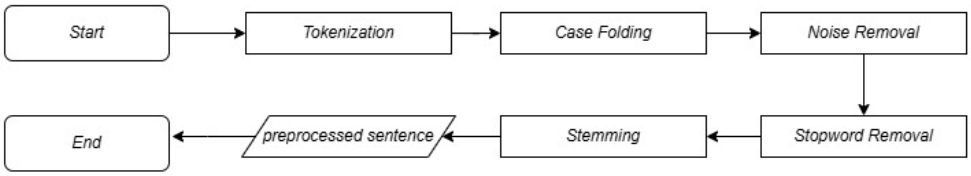
			'enggak', 'tahu', ',', 'belum', 'sempat', 'jenguk', 'dan', 'enggak', 'selamanya', 'bisa', 'dijenguk', 'juga', '.'], ['Enggak', 'tahu', 'berat', 'sekali', 'apa', 'bagaimana', ',', "", 'tutur', 'Ryan', '.']], [['Walau', 'sudah', 'setahun', 'menderita', 'sakit', ',', 'Lula', 'tak', 'mengetahui', 'apa', 'penyebab', 'pasti', 'kematian', 'Dr', 'Oz', 'Indonesia', 'itu', '.'], ['Meski', 'demikian', ',', 'ia', 'mendengar', 'beberapa', 'kabar', 'yang', 'menyebut', 'bahwa', 'penyebab', 'Ryan', 'meninggal', 'adalah', 'karena', 'jatuh', 'di', 'kamar', 'mandi', '.']], [["“", 'Saya', 'tidak', 'tahu', ',', 'barangkali', 'penyakit', 'yang', 'dulu', 'sama', 'yang', 'sekarang', 'berbeda', ',', 'atau', 'penyebab', 'kematiannya', 'beda', 'dari', 'penyakit', 'sebelumnya', '.'], ['Kita', 'kan', 'enggak', 'bisa', 'mengambil', 'kesimpulan', ',', "", 'kata', 'Lula', '.']], [['Ryan', 'Thamrin', 'terkenal', 'sebagai', 'dokter', 'yang', 'rutin', 'membagikan', 'tips', 'dan', 'informasi', 'kesehatan', 'lewat', 'tayangan', 'Dokter', 'Oz', 'Indonesia', '.']], [['Ryan', 'menempuh', 'Pendidikan', 'Dokter', 'pada',			

			'tahun', '2002', 'di', 'Fakultas', 'Kedokteran', 'Universitas', 'Gadjah', 'Mada', '.'], ['Dia', 'kemudian', 'melanjutkan', 'pendidikan', 'Klinis', 'Kesehatan', 'Reproduksi', 'dan', 'Penyakit', 'Menular', 'Seksual', 'di', 'Mahachulalongkornrajavidyalaya', 'University', ',', 'Bangkok', ',', 'Thailand', 'pada', '2004', '.']]]			

3.3 *Preprocessing*

Preprocessing merupakan langkah selanjutnya dalam penelitian ini yang bertujuan untuk mempersiapkan teks agar dapat diolah secara lebih efektif oleh model serta menyempurnakan struktur *input*an. Tahapan ini dimulai dengan *tokenization* yang berfungsi memecah masukan berupa kalimat ke dalam unit kata (token) individual, kemudian dilanjutkan dengan *case folding* untuk mengkonversi semua teks ke huruf kecil guna menyamakan format penulisan. Selanjutnya dilakukan *noise removal* untuk mengeliminasi komponen seperti tanda baca dan simbol yang tidak diperlukan. Kemudian diterapkan *stopword removal* untuk membuang kata-kata umum seperti "dan" atau "yang" yang tidak memberikan nilai signifikan terhadap makna. Tahap akhir yaitu *stemming*, yang mentransformasi kata-kata ke bentuk dasarnya menggunakan library *Sastrawi* untuk menyatukan kata-kata yang memiliki makna sejenis.Setiap langkah ini bertujuan menyederhanakan *token* tanpa menghilangkan informasi penting, sehingga *input*

lebih terstruktur dan siap diolah oleh model. Untuk penggambaran diagram *preprocessing* dari tahap ini terdapat pada Gambar 3.3



Gambar 3.3 *Flowchart Preprocessing*

3.3.1 *Tokenization*

Tahapan *tokenization* merupakan langkah awal dalam *preprocessing* yang berfungsi memecah *input* kalimat menjadi unit-unit kata individual yang disebut *token*. Pada proses ini, kalimat utuh dipisahkan berdasarkan spasi sehingga setiap kata dapat diproses secara terpisah. Misalnya, kalimat "UIN Maulana Malik Ibrahim Malang" akan dipecah menjadi lima token terpisah: "UIN", "Maulana", "Malik", "Ibrahim", "Malang". *Tokenization* menjadi fundamental dalam analisis teks karena memungkinkan sistem untuk mengolah setiap kata secara independen pada tahapan *preprocessing* selanjutnya seperti *case folding*, *noise removal*, *stopword removal*, dan *stemming*. Dengan memisahkan kalimat menjadi *token-token*, sistem dapat melakukan perhitungan frekuensi kata dan analisis linguistik yang lebih terperinci untuk proses peringkasan dokumen. Pada Tabel 3.2 adalah contoh hasil sebelum dan sesudah dilakukan *tokenization*

Tabel 3.2 Contoh Sebelum dan Sesudah *Tokenization*

Id	**Sebelum *Tokenization***	**Sesudah *Tokenization***
0	Jakarta, CNN Indonesia - - Dokter Ryan Thamrin, yang terkenal lewat acara Dokter	['Jakarta', ',', 'CNN', 'Indonesia', 'Dokter', 'Ryan', 'Thamrin', ',', 'yang', 'terkenal', 'lewat', 'acara',

	Oz Indonesia, meninggal dunia pada Jumat (4/8) dini hari.	'Dokter', 'Oz', 'Indonesia', ',', 'meninggal', 'dunia', 'pada', 'Jumat', '(', ')', 'dini', 'hari', '.']
1	Dokter Lula Kamal yang merupakan selebriti sekaligus rekan kerja Ryan menyebut kawannya itu sudah sakit sejak setahun yang lalu.	['Dokter', 'Lula', 'Kamal', 'yang', 'merupakan', 'selebriti', 'sekaligus', 'rekan', 'kerja', 'Ryan', 'menyebut', 'kawannya', 'itu', 'sudah', 'sakit', 'sejak', 'setahun', 'yang', 'lalu', '.']
2	Lula menuturkan, sakit itu membuat Ryan mesti vakum dari semua kegiatannya, termasuk menjadi pembawa acara Dokter Oz Indonesia.	['Lula', 'menuturkan', ',', 'sakit', 'itu', 'membuat', 'Ryan', 'mesti', 'vakum', 'dari', 'semua', 'kegiatannya', ',', 'termasuk', 'menjadi', 'pembawa', 'acara', 'Dokter', 'Oz', 'Indonesia', '.']
3	Kondisi itu membuat Ryan harus kembali ke kampung halamannya di Pekanbaru, Riau untuk menjalani istirahat.	['Kondisi', 'itu', 'membuat', 'Ryan', 'harus', 'kembali', 'ke', 'kampung', 'halamannya', 'di', 'Pekanbaru', ',', 'Riau', 'untuk', 'menjalani', 'istirahat', '.']
…	…	…
15	Ryan Thamrin terkenal sebagai dokter yang rutin membagikan tips dan informasi kesehatan lewat tayangan Dokter Oz Indonesia.	['Ryan', 'Thamrin', 'terkenal', 'sebagai', 'dokter', 'yang', 'rutin', 'membagikan', 'tips', 'dan', 'informasi', 'kesehatan', 'lewat', 'tayangan', 'Dokter', 'Oz', 'Indonesia', '.']
16	Ryan menempuh Pendidikan Dokter pada tahun 2002 di Fakultas Kedokteran Universitas Gadjah Mada.	['Ryan', 'menempuh', 'Pendidikan', 'Dokter', 'pada', 'tahun', 'di', 'Fakultas', 'Kedokteran', 'Universitas', 'Gadjah', 'Mada', '.']
17	Dia kemudian melanjutkan pendidikan Klinis Kesehatan Reproduksi dan Penyakit Menular Seksual di Mahachulalongkornrajavidyalaya University, Bangkok, Thailand pada 2004.	['Dia', 'kemudian', 'melanjutkan', 'pendidikan', 'Klinis', 'Kesehatan', 'Reproduksi', 'dan', 'Penyakit', 'Menular', 'Seksual', 'di', 'Mahachulalongkornrajavidyalaya', 'University', ',', 'Bangkok', ',', 'Thailand', 'pada', '.']

3.3.2 *Case Folding*

Tahapan selanjutnya adalah *case folding* yang akan mengubah seluruh kata yang menggunakan huruf kapital menjadi huruf kecil. Tujuan dari *case folding* adalah untuk menghindari perbedaan makna yang dapat terjadi ketika adanya huruf kapital suatu kata. Sebagai contoh, kata "Jakarta" dan "jakarta" seharusnya dianggap sama dalam pemrosesan teks, meskipun satu menggunakan huruf kapital di awal dan yang lainnya tidak. Dengan melakukan *case folding*, semua huruf dalam teks diubah menjadi huruf kecil, sehingga mengurangi kompleksitas analisis dan memastikan konsistensi dalam interpretasi kata-kata di seluruh teks. Pada Tabel 3.3 adalah contoh hasil sebelum dan sesudah dilakukan *case folding*

Tabel 3.3 Contoh Sebelum dan Sesudah *Case Folding*

Id	**Sebelum *Case Folding***	**Sesudah *Case Folding***
0	['Jakarta', ',', 'CNN', 'Indonesia', 'Dokter', 'Ryan', 'Thamrin', ',', 'yang', 'terkenal', 'lewat', 'acara', 'Dokter', 'Oz', 'Indonesia', ',', 'meninggal', 'dunia', 'pada', 'Jumat', '(', ')', 'dini', 'hari', '.']	['jakarta', ',', 'cnn', 'indonesia', 'dokter', 'ryan', 'thamrin', ',', 'yang', 'terkenal', 'lewat', 'acara', 'dokter', 'oz', 'indonesia', ',', 'meninggal', 'dunia', 'pada', 'jumat', '(', ')', 'dini', 'hari', '.']
1	['Dokter', 'Lula', 'Kamal', 'yang', 'merupakan', 'selebriti', 'sekaligus', 'rekan', 'kerja', 'Ryan', 'menyebut', 'kawannya', 'itu', 'sudah', 'sakit', 'sejak', 'setahun', 'yang', 'lalu', '.']	['dokter', 'lula', 'kamal', 'yang', 'merupakan', 'selebriti', 'sekaligus', 'rekan', 'kerja', 'ryan', 'menyebut', 'kawannya', 'itu', 'sudah', 'sakit', 'sejak', 'setahun', 'yang', 'lalu', '.']
2	['Lula', 'menuturkan', ',', 'sakit', 'itu', 'membuat', 'Ryan', 'mesti', 'vakum', 'dari', 'semua', 'kegiatannya', ',', 'termasuk', 'menjadi', 'pembawa', 'acara', 'Dokter', 'Oz', 'Indonesia', '.']	['lula', 'menuturkan', ',', 'sakit', 'itu', 'membuat', 'ryan', 'mesti', 'vakum', 'dari', 'semua', 'kegiatannya', ',', 'termasuk', 'menjadi', 'pembawa', 'acara', 'dokter', 'oz', 'indonesia', '.']
3	['Kondisi', 'itu', 'membuat', 'Ryan', 'harus', 'kembali', 'ke', 'kampung', 'halamannya',	['kondisi', 'itu', 'membuat', 'ryan', 'harus', 'kembali', 'ke', 'kampung', 'halamannya', 'di', 'pekanbaru', ',',

	'di', 'Pekanbaru', ',', 'Riau', 'untuk', 'menjalani', 'istirahat', '.']	'riau', 'untuk', 'menjalani', 'istirahat', '.']
…	…	…
15	['Ryan', 'Thamrin', 'terkenal', 'sebagai', 'dokter', 'yang', 'rutin', 'membagikan', 'tips', 'dan', 'informasi', 'kesehatan', 'lewat', 'tayangan', 'Dokter', 'Oz', 'Indonesia', '.']	['ryan', 'thamrin', 'terkenal', 'sebagai', 'dokter', 'yang', 'rutin', 'membagikan', 'tips', 'dan', 'informasi', 'kesehatan', 'lewat', 'tayangan', 'dokter', 'oz', 'indonesia', '.']
16	['Ryan', 'menempuh', 'Pendidikan', 'Dokter', 'pada', 'tahun', 'di', 'Fakultas', 'Kedokteran', 'Universitas', 'Gadjah', 'Mada', '.']	['ryan', 'menempuh', 'pendidikan', 'dokter', 'pada', 'tahun', 'di', 'fakultas', 'kedokteran', 'universitas', 'gadjah', 'mada', '.']
17	['Dia', 'kemudian', 'melanjutkan', 'pendidikan', 'Klinis', 'Kesehatan', 'Reproduksi', 'dan', 'Penyakit', 'Menular', 'Seksual', 'di', 'Mahachulalongkornrajavidyalaya', 'University', ',', 'Bangkok', ',', 'Thailand', 'pada', '.']	['dia', 'kemudian', 'melanjutkan', 'pendidikan', 'klinis', 'kesehatan', 'reproduksi', 'dan', 'penyakit', 'menular', 'seksual', 'di', 'mahachulalongkornrajavidyalaya', 'university', ',', 'bangkok', ',', 'thailand', 'pada', '.']

3.3.3 *Noise Removal*

Langkah selanjutnya adalah *noise removal*. *Noise removal* adalah proses membersihkan teks dari elemen-elemen yang tidak relevan atau tidak memiliki nilai informasi dalam konteks analisis. Elemen ini meliputi tanda baca, simbol dan karakter. Dengan menghilangkan *noise*, teks menjadi lebih bersih dan fokus pada konten yang relevan untuk analisis. Pada Tabel 3.4 adalah contoh hasil sebelum dan sesudah *noise removal*

Tabel 3.4 Contoh Sebelum dan Sesudah *Noise Removal*

Id	**Sebelum *Noise Removal***	**Sesudah *Noise Removal***

0	['jakarta', ',', 'cnn', 'indonesia', 'dokter', 'ryan', 'thamrin', ',', 'yang', 'terkenal', 'lewat', 'acara', 'dokter', 'oz', 'indonesia', ',', 'meninggal', 'dunia', 'pada', 'jumat', '(', ')', 'dini', 'hari', '.']	['jakarta', 'cnn', 'indonesia', 'dokter', 'ryan', 'thamrin', 'yang', 'terkenal', 'lewat', 'acara', 'dokter', 'oz', 'indonesia', 'meninggal', 'dunia', 'pada', 'jumat', 'dini', 'hari']
1	['dokter', 'lula', 'kamal', 'yang', 'merupakan', 'selebriti', 'sekaligus', 'rekan', 'kerja', 'ryan', 'menyebut', 'kawannya', 'itu', 'sudah', 'sakit', 'sejak', 'setahun', 'yang', 'lalu', '.']	['dokter', 'lula', 'kamal', 'yang', 'merupakan', 'selebriti', 'sekaligus', 'rekan', 'kerja', 'ryan', 'menyebut', 'kawannya', 'itu', 'sudah', 'sakit', 'sejak', 'setahun', 'yang', 'lalu']
2	['lula', 'menuturkan', ',', 'sakit', 'itu', 'membuat', 'ryan', 'mesti', 'vakum', 'dari', 'semua', 'kegiatannya', ',', 'termasuk', 'menjadi', 'pembawa', 'acara', 'dokter', 'oz', 'indonesia', '.']	['lula', 'menuturkan', 'sakit', 'itu', 'membuat', 'ryan', 'mesti', 'vakum', 'dari', 'semua', 'kegiatannya', 'termasuk', 'menjadi', 'pembawa', 'acara', 'dokter', 'oz', 'indonesia']
3	['kondisi', 'itu', 'membuat', 'ryan', 'harus', 'kembali', 'ke', 'kampung', 'halamannya', 'di', 'pekanbaru', ',', 'riau', 'untuk', 'menjalani', 'istirahat', '.']	['kondisi', 'itu', 'membuat', 'ryan', 'harus', 'kembali', 'ke', 'kampung', 'halamannya', 'di', 'pekanbaru', 'riau', 'untuk', 'menjalani', 'istirahat']
…	…	…
15	['ryan', 'thamrin', 'terkenal', 'sebagai', 'dokter', 'yang', 'rutin', 'membagikan', 'tips', 'dan', 'informasi', 'kesehatan', 'lewat', 'tayangan', 'dokter', 'oz', 'indonesia', '.']	['ryan', 'thamrin', 'terkenal', 'sebagai', 'dokter', 'yang', 'rutin', 'membagikan', 'tips', 'dan', 'informasi', 'kesehatan', 'lewat', 'tayangan', 'dokter', 'oz', 'indonesia']
16	['ryan', 'menempuh', 'pendidikan', 'dokter', 'pada', 'tahun', 'di', 'fakultas', 'kedokteran', 'universitas', 'gadjah', 'mada', '.']	['ryan', 'menempuh', 'pendidikan', 'dokter', 'pada', 'tahun', 'di', 'fakultas', 'kedokteran', 'universitas', 'gadjah', 'mada']
17	['dia', 'kemudian', 'melanjutkan', 'pendidikan', 'klinis', 'kesehatan', 'reproduksi', 'dan', 'penyakit', 'menular', 'seksual', 'di', 'mahachulalongkornrajavidyalaya', 'university', ',', 'bangkok', ',', 'thailand', 'pada', '.']	['dia', 'kemudian', 'melanjutkan', 'pendidikan', 'klinis', 'kesehatan', 'reproduksi', 'dan', 'penyakit', 'menular', 'seksual', 'di', 'mahachulalongkornrajavidyalaya', 'university', 'bangkok', 'thailand', 'pada']

3.3.4 *Stopword Removal*

Tahap selanjutnya adalah *stopword removal* merupakan proses menghapus kata-kata umum yang tidak memiliki kontribusi signifikan terhadap makna teks, seperti "pada", "di", "yang" dan sebagainya. *Stopword* tidak memberikan informasi penting untuk analisis topik utama dalam dokumen, sehingga penghapusannya membantu mengurangi dimensi data tanpa kehilangan informasi utama. Tabel 3.5 memperlihatkan contoh sebelum dan sesudah proses *stopword removal*

Tabel 3.5 Contoh Sebelum dan Sesudah *Stopword Removal*

Id	**Sebelum *Stopword Removal***	**Sesudah *Stopword Removal***
0	['jakarta', 'cnn', 'indonesia', 'dokter', 'ryan', 'thamrin', 'yang', 'terkenal', 'lewat', 'acara', 'dokter', 'oz', 'indonesia', 'meninggal', 'dunia', 'pada', 'jumat', 'dini', 'hari']	['jakarta', 'cnn', 'indonesia', 'dokter', 'ryan', 'thamrin', 'terkenal', 'lewat', 'acara', 'dokter', 'oz', 'indonesia', 'meninggal', 'dunia', 'jumat', 'dini', 'hari']
1	['dokter', 'lula', 'kamal', 'yang', 'merupakan', 'selebriti', 'sekaligus', 'rekan', 'kerja', 'ryan', 'menyebut', 'kawannya', 'itu', 'sudah', 'sakit', 'sejak', 'setahun', 'yang', 'lalu']	['dokter', 'lula', 'kamal', 'merupakan', 'selebriti', 'sekaligus', 'rekan', 'kerja', 'ryan', 'menyebut', 'kawannya', 'sakit', 'sejak', 'setahun', 'lalu']
2	['lula', 'menuturkan', 'sakit', 'itu', 'membuat', 'ryan', 'mesti', 'vakum', 'dari', 'semua', 'kegiatannya', 'termasuk', 'menjadi', 'pembawa', 'acara', 'dokter', 'oz', 'indonesia']	['lula', 'menuturkan', 'sakit', 'membuat', 'ryan', 'mesti', 'vakum', 'semua', 'kegiatannya', 'termasuk', 'menjadi', 'pembawa', 'acara', 'dokter', 'oz', 'indonesia']
3	['kondisi', 'itu', 'membuat', 'ryan', 'harus', 'kembali', 'ke', 'kampung', 'halamannya', 'di', 'pekanbaru', 'riau', 'untuk', 'menjalani', 'istirahat']	['kondisi', 'membuat', 'ryan', 'kampung', 'halamannya', 'pekanbaru', 'riau', 'menjalani', 'istirahat']
…	…	…
15	['ryan', 'thamrin', 'terkenal', 'sebagai', 'dokter', 'yang', 'rutin', 'membagikan', 'tips', 'dan', 'informasi', 'kesehatan',	['ryan', 'thamrin', 'terkenal', 'dokter', 'rutin', 'membagikan', 'tips',

	'lewat', 'tayangan', 'dokter', 'oz', 'indonesia']	'informasi', 'kesehatan', 'lewat', 'tayangan', 'dokter', 'oz', 'indonesia']
16	['ryan', 'menempuh', 'pendidikan', 'dokter', 'pada', 'tahun', 'di', 'fakultas', 'kedokteran', 'universitas', 'gadjah', 'mada']	['ryan', 'menempuh', 'pendidikan', 'dokter', 'tahun', 'fakultas', 'kedokteran', 'universitas', 'gadjah', 'mada']
17	['dia', 'kemudian', 'melanjutkan', 'pendidikan', 'klinis', 'kesehatan', 'reproduksi', 'dan', 'penyakit', 'menular', 'seksual', 'di', 'mahachulalongkornrajavidyalaya', 'university', 'bangkok', 'thailand', 'pada']	['kemudian', 'melanjutkan', 'pendidikan', 'klinis', 'kesehatan', 'reproduksi', 'penyakit', 'menular', 'seksual', 'mahachulalongkornrajavidyalaya', 'university', 'bangkok', 'thailand']

3.3.4 *Stemming*

Stemming adalah proses mengubah kata-kata menjadi kata dasarnya. Adapun tujuan dari *stemming* adalah menghilangkan infleksi sehingga kata yang memiliki akar sama akan mudah direpresentasikan dengan seragam. Setiap *token* akan diperiksa dalam kamus kata dasar. Jika *token* tidak ditemukan dalam kamus tersebut maka *token* dianggap sebagai kata berimbuhan. Selanjutnya, proses *stemming* dilakukan dengan menghapus akhiran, imbuhan turunan dan penghapusan imbuhan awal. Tabel 3.6 memperlihatkan contoh sebelum dan sesudah proses *stemming*

Tabel 3.6 Contoh Sebelum dan Sesudah *Stemming*

Id	**Sebelum *Stemming***	**Sesudah *Stemming***
0	['jakarta', 'cnn', 'indonesia', 'dokter', 'ryan', 'thamrin', 'terkenal', 'lewat', 'acara', 'dokter', 'oz', 'indonesia', 'meninggal', 'dunia', 'jumat', 'dini', 'hari']	['jakarta', 'cnn', 'indonesia', 'dokter', 'ryan', 'thamrin', 'kenal', 'lewat', 'acara', 'dokter', 'oz', 'indonesia', 'tinggal', 'dunia', 'jumat', 'dini', 'hari']

1	['dokter', 'lula', 'kamal', 'merupakan', 'selebriti', 'sekaligus', 'rekan', 'kerja', 'ryan', 'menyebut', 'kawannya', 'sakit', 'sejak', 'setahun', 'lalu']	['dokter', 'lula', 'kamal', 'rupa', 'selebriti', 'sekaligus', 'rekan', 'kerja', 'ryan', 'sebut', 'kawan', 'sakit', 'sejak', 'tahun', 'lalu']
2	['lula', 'menuturkan', 'sakit', 'membuat', 'ryan', 'mesti', 'vakum', 'semua', 'kegiatannya', 'termasuk', 'menjadi', 'pembawa', 'acara', 'dokter', 'oz', 'indonesia']	['lula', 'tutur', 'sakit', 'buat', 'ryan', 'mesti', 'vakum', 'semua', 'giat', 'masuk', 'jadi', 'bawa', 'acara', 'dokter', 'oz', 'indonesia']
3	['kondisi', 'membuat', 'ryan', 'kampung', 'halamannya', 'pekanbaru', 'riau', 'menjalani', 'istirahat']	['kondisi', 'buat', 'ryan', 'kampung', 'halaman', 'pekanbaru', 'riau', 'jalan', 'istirahat']
…	…	…
15	['ryan', 'thamrin', 'terkenal', 'dokter', 'rutin', 'membagikan', 'tips', 'informasi', 'kesehatan', 'lewat', 'tayangan', 'dokter', 'oz', 'indonesia']	['ryan', 'thamrin', 'kenal', 'dokter', 'rutin', 'bagi', 'tips', 'informasi', 'sehat', 'lewat', 'tayang', 'dokter', 'oz', 'indonesia']
16	['ryan', 'menempuh', 'pendidikan', 'dokter', 'tahun', 'fakultas', 'kedokteran', 'universitas', 'gadjah', 'mada']	['ryan', 'tempuh', 'didik', 'dokter', 'tahun', 'fakultas', 'dokter', 'universitas', 'gadjah', 'mada']
17	['kemudian', 'melanjutkan', 'pendidikan', 'klinis', 'kesehatan', 'reproduksi', 'penyakit', 'menular', 'seksual', 'mahachulalongkornrajavidyalaya', 'university', 'bangkok', 'thailand']	['kemudian', 'lanjut', 'didik', 'klinis', 'sehat', 'reproduksi', 'sakit', 'tular', 'seksual', 'mahachulalongkornrajavidyalaya', 'university', 'bangkok', 'thailand']

3.4 Pembobotan TF-IDF

Pada tahap ini, sistem menghitung nilai TF-IDF untuk setiap kata dalam kalimat. *Term Frequency* (TF) mengukur seberapa sering kata muncul dalam kalimat, dihitung dengan menghitung frekuensi kemunculan kata tersebut pada tiap kalimat. Sedangkan *Inverse Document Frequency* (IDF) mengukur seberapa unik kata tersebut di seluruh kalimat, dihitung menggunakan Persamaan 3.1.

$$IDF(t) = log(\frac{N}{DF(t)}) \quad (3.1)$$

Keterangan:
$IDF(t)$ = nilai *Inverse Document Frequency* untuk *term* t
N = jumlah total kalimat dalam dokumen
$DF(t)$ = jumlah kalimat yang mengandung *term* t

Nilai TF-IDF merupakan hasil perkalian dari kedua nilai tersebut, yang dapat dilihat pada Persamaan 3.2

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \quad (3.2)$$

Keterangan:
$TF - IDF(t,d)$ = bobot TF-IDF untuk term t dalam kalimat d
$TF(t,d)$ = frekuensi kemunculan *term* t dalam kalimat d
$IDF(t)$ = nilai *Inverse Document Frequency* untuk *term* t

Nilai TF-IDF yang tinggi menunjukkan bahwa kata tersebut penting untuk kalimat tertentu, tetapi tidak umum di seluruh koleksi kalimat dalam dokumen. Tabel 3.7 memperlihatkan proses perhitungan bobot TF-IDF

Tabel 3.7 Perhitungan Bobot TF-IDF

Kalimat	**Kata**	**TF**	**N**	**DF**	**IDF**	**TF-IDF**
jakarta cnn indonesia dokter ryan thamrin kenal lewat acara dokter oz indonesia tinggal dunia jumat dini hari	dokter	2	18	5	1.280934	2.561868
	indonesia	2		4	1.504077	3.008155
	acara	1		2	2.197225	2.197225
	cnn	1		1	2.890372	2.890372
	…	…		…	…	…
	ryan	1		10	0.587787	0.587787

	thamrin	1		2	2.197225	2.197225
	tinggal	1		2	2.197225	2.197225
dokter lula kamal rupa selebriti sekaligus rekan kerja ryan sebut kawan sakit sejak tahun lalu	dokter	1		5	1.280934	1.280934
	lula	1		6	1.098612	1.098612
	kamal	1		1	2.890372	2.890372
	rupa	1		1	2.890372	2.890372
	…	…		…	…	…
	sejak	1		2	2.197225	2.197225
	tahun	1	18	4	1.504077	1.504077
	lalu	1		2	2.197225	2.197225
…	…	…		…	…	…
ryan thamrin kenal dokter rutin bagi tips informasi sehat lewat tayang dokter oz indonesia	dokter	2		5	1.280934	2.561868
	bagi	1		1	2.890372	2.890372
	indonesia	1		4	1.504077	1.504077
	informasi	1		1	2.890372	2.890372
	…	…		…	…	…
	kenal	1		3	1.791759	1.791759

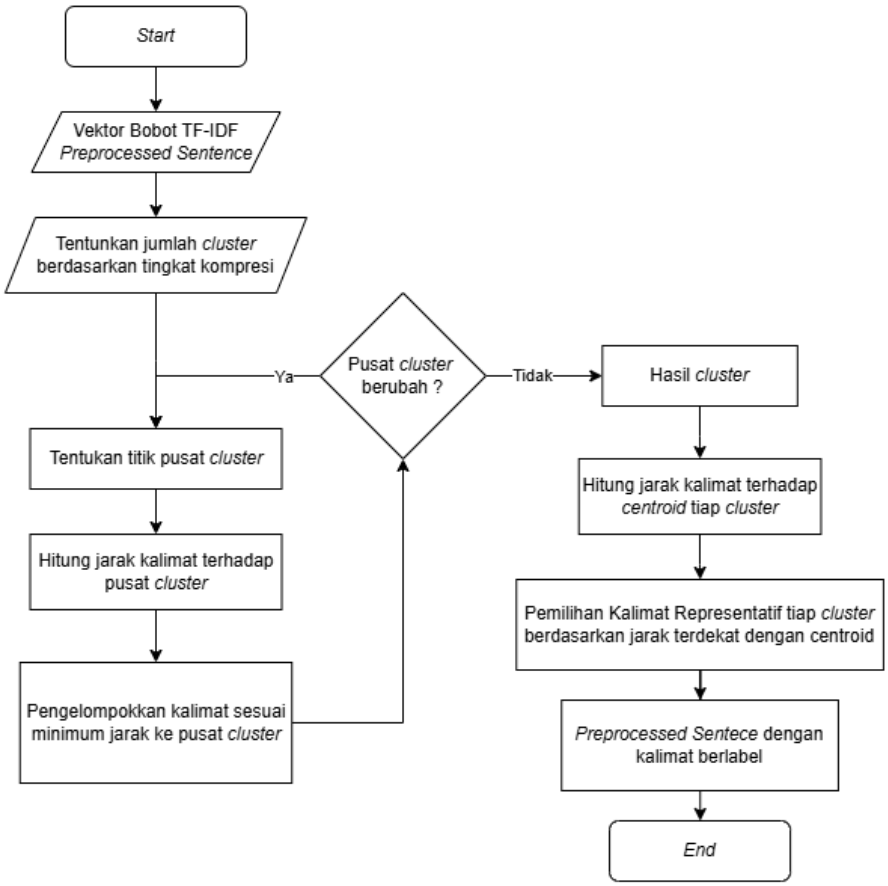
	lewat	1		2	2.197225	2.197225
	oz	1		4	1.504077	1.504077
ryan tempuh didik dokter tahun fakultas dokter universitas gadjah mada	dokter	2		5	1.280934	2.561868
	didik	1		2	2.197225	2.197225
	fakultas	1		1	2.890372	2.890372
	…	…		…	…	…
	gadjah	1		1	2.890372	2.890372
	mada	1	18	1	2.890372	2.890372
	tempuh	1		1	2.890372	2.890372
kemudian lanjut didik klinis sehat reproduksi sakit tular seksual mahachulalongkorn rajavidyalaya university bangkok thailand	bangkok	1		1	2.890372	2.890372
	klinis	1		1	2.890372	2.890372
	sakit	1		8	0.810930	0.810930
	…	…		…	…	…
	reproduksi	1		1	2.890372	2.890372
	tular	1		1	2.890372	2.890372

3.5 *K-Means Clustering*

Dalam proses peringkasan dokumen, tahap *K-Means Clustering* digunakan untuk mengelompokkan kalimat-kalimat berdasarkan kemiripan semantiknya.

Seperti ditunjukkan pada Gambar 3.4, proses ini dimulai dengan vektor bobot TF-IDF dari kalimat yang telah di *preprocessing*. Jumlah *cluster* (*k*) ditentukan melalui proses pencarian *k* optimal yang mempertimbangkan kualitas *clustering*.

Setelah menentukan jumlah *cluster*, algoritma akan menentukan titik pusat awal untuk setiap *cluster*. Kemudian, jarak setiap kalimat terhadap pusat *cluster* dihitung, dan kalimat dikelompokkan berdasarkan jarak terdekat ke pusat *cluster* tersebut. Proses ini diulang dengan memperbarui pusat *cluster* hingga tidak ada perubahan pada pusat *cluster* (konvergen). Setelah konvergen, algoritma menghitung jarak setiap kalimat terhadap *centroid* final dari masing-masing *cluster*. Kalimat dengan jarak terdekat ke *centroid* pada tiap *cluster* dipilih sebagai kalimat representatif untuk *cluster* tersebut. Hasil akhirnya adalah vektor bobot TF-IDF dari kalimat-kalimat terpilih yang telah diberi label yang siap untuk dijadikan ringkasan akhir. Pendekatan ini memastikan bahwa jumlah kalimat dalam ringkasan akhir sesuai dengan tingkat kompresi yang diinginkan, sekaligus mempertahankan kalimat-kalimat yang paling representatif dari dokumen asli.



Gambar 3.4 *Flowchart K-Means Clustering*

Pada Gambar 3.4 tersebut memberikan tahapan-tahapan penting mulai dari *input* vektor bobot TF-IDF hingga pemilihan kalimat representatif untuk ringkasan akhir. Untuk memahami proses ini secara lebih detail, berikut akan dijelaskan setiap tahapan dalam algoritma *K-Means Clustering*.

3.5.1 Data Input

Tahap pertama dalam algoritma *K-Means Clustering* adalah mempersiapkan data *input* berupa vektor bobot TF-IDF dari kalimat-kalimat yang telah di*preprocessing*. Setiap kalimat direpresentasikan sebagai vektor numerik

yang mencerminkan frekuensi dan signifikansi kata-kata di dalamnya. Vektor bobot TF-IDF dapat dilihat pada Gambar 3.5

```
[DEBUG] Contoh Matriks Bobot TF-IDF:
Kalimat 1 (ID: 0): [2.1972, 0.0000, 0.0000, 0.0000, 0.0000, ... 0.0000, 0.0000, 0.0000, 0.0000, 0.0000]
Kalimat 2 (ID: 1): [0.0000, 0.0000, 0.0000, 0.0000, 0.0000, ... 0.0000, 0.0000, 0.0000, 0.0000, 0.0000]
Kalimat 3 (ID: 2): [2.1972, 0.0000, 0.0000, 0.0000, 0.0000, ... 0.0000, 2.1972, 0.0000, 0.0000, 2.8904]
Kalimat 4 (ID: 3): [0.0000, 0.0000, 0.0000, 0.0000, 0.0000, ... 0.0000, 0.0000, 0.0000, 0.0000, 0.0000]
Kalimat 5 (ID: 4): [0.0000, 0.0000, 0.0000, 0.0000, 0.0000, ... 0.0000, 0.0000, 0.0000, 0.0000, 0.0000]
...
Kalimat 14 (ID: 13): [0.0000, 0.0000, 0.0000, 0.0000, 0.0000, ... 0.0000, 0.0000, 0.0000, 0.0000, 0.0000]
Kalimat 15 (ID: 14): [0.0000, 0.0000, 0.0000, 2.8904, 0.0000, ... 0.0000, 0.0000, 0.0000, 0.0000, 0.0000]
Kalimat 16 (ID: 15): [0.0000, 0.0000, 0.0000, 0.0000, 0.0000, ... 0.0000, 0.0000, 0.0000, 0.0000, 0.0000]
Kalimat 17 (ID: 16): [0.0000, 0.0000, 0.0000, 0.0000, 0.0000, ... 0.0000, 0.0000, 2.8904, 0.0000, 0.0000]
Kalimat 18 (ID: 17): [0.0000, 0.0000, 0.0000, 0.0000, 0.0000, ... 2.8904, 0.0000, 0.0000, 2.8904, 0.0000]
```

Gambar 3.5 Vektor Bobot TF-IDF

3.5.2 Inisiasi *Centroid* Awal

Pada tahap ini, sistem menginisialisasi titik pusat awal (*centroid*) untuk setiap *cluster* yang telah ditentukan. Jumlah *k* ditentukan melalui pencarian *k* optimal. Untuk percobaan ini, algoritma menggunakan K=2. Titik-titik pusat ini berfungsi sebagai acuan awal untuk mengelompokkan kalimat-kalimat berdasarkan kemiripan semantiknya.

3.5.3 Menghitung Jarak *Centroid*

Setelah centroid awal ditetapkan, sistem menghitung jarak setiap kalimat terhadap semua *centroid* menggunakan metrik jarak *Euclidean* seperti pada Persamaan 2.1. Jarak ini menjadi ukuran kemiripan antara kalimat dengan pusat *cluster*. Semakin kecil jarak antara kalimat dengan suatu *centroid*, semakin tinggi kemiripan semantiknya dengan *cluster* tersebut. Setiap kalimat kemudian ditetapkan ke *cluster* dengan jarak *centroid* terdekat yang dapat dilihat pada Tabel 3.8

Tabel 3.8 Pengelompokan Kalimat ke *Cluster*

Cluster	**Id**	**Kalimat**	**Jarak ke *centroid***
1	1	dokter lula kamal rupa selebriti sekaligus rekan kerja ryan sebut kawan sakit sejak tahun lalu	8.339108
	2	lula tutur sakit buat ryan mesti vakum semua giat masuk jadi bawa acara dokter oz indonesia	8.289591
	3	kondisi buat ryan kampung halaman pekanbaru riau jalan istirahat	7.262901
	…	…	…
	16	ryan tempuh didik dokter tahun fakultas dokter universitas gadjah mada	7.180636
	17	kemudian lanjut didik klinis sehat reproduksi sakit tular seksual mahachulalongkornrajavidyalaya university bangkok thailand	9.188398
2	0	jakarta cnn indonesia dokter ryan thamrin kenal lewat acara dokter oz indonesia tinggal dunia jumat dini hari	4.999301
	15	ryan thamrin kenal dokter rutin bagi tips informasi sehat lewat tayang dokter oz indonesia	4.999301

3.5.4 Memperbarui *Centroid*

Tahap ini melibatkan perhitungan ulang pusat *cluster* (*centroid*) berdasarkan kalimat-kalimat yang telah dikelompokkan pada inisiasi sebelumnya

menggunakan Persamaan 2.2. *Centroid* baru dihitung sebagai rata-rata vektor dari semua kalimat dalam *cluster* tersebut. Setelah *centroid* diperbarui, sistem kembali menghitung jarak dan mengelompokkan kalimat. Proses ini berulang hingga tidak ada perubahan yang signifikan pada posisi *centroid* (konvergen) atau jumlah iterasi maksimum tercapai.

3.5.5 Hasil *Cluster*

Setelah algoritma konvergen, diperoleh hasil pengelompokan final dimana setiap kalimat telah ditetapkan ke salah satu dari 2 *cluster*. Setiap *cluster* berisi kalimat-kalimat dengan karakteristik semantik yang mirip berdasarkan representasi vektor TF-IDF mereka. Hasil pengelompokan ini mencerminkan struktur semantik dari dokumen asli yang telah dipartisi ke dalam *k* kelompok berbeda.

3.5.6 Memilih Kalimat Representatif

Pada tahap akhir, sistem memilih satu kalimat representatif dari setiap *cluster* untuk dijadikan bagian dari ringkasan. Pemilihan dilakukan dengan menghitung jarak setiap kalimat dalam *cluster* terhadap *centroid* menggunakan *Euclidean distance* seperti pada Persamaan 2.1 lalu memilih kalimat dengan jarak terdekat ke *centroid*. Dapat dilihat pada Tabel 3.9, kalimat-kalimat terpilih ini dianggap paling mewakili isi dari masing-masing *cluster* dan bersama-sama membentuk ringkasan yang komprehensif dari dokumen asli.

Tabel 3.9 Kalimat Representatif dari tiap *Cluster*

Cluster	**Id**	**Kalimat**	**Jarak ke *centroid***

1	8	dokter lula kamal rupa selebriti sekaligus rekan kerja ryan sebut kawan sakit sejak tahun lalu	3.847737
	4	lula tutur sakit buat ryan mesti vakum semua giat masuk jadi bawa acara dokter oz indonesia	4.867305
	11	kondisi buat ryan kampung halaman pekanbaru riau jalan istirahat	5.710676
	…	…	…
	5	ryan tempuh didik dokter tahun fakultas dokter universitas gadjah mada	6.117109
	9	kemudian lanjut didik klinis sehat reproduksi sakit tular seksual mahachulalongkornrajavidyalaya university bangkok thailand	6.475695
2	0	jakarta cnn indonesia dokter ryan thamrin kenal lewat acara dokter oz indonesia tinggal dunia jumat dini hari	4.999301
	15	ryan thamrin kenal dokter rutin bagi tips informasi sehat lewat tayang dokter oz indonesia	4.999301

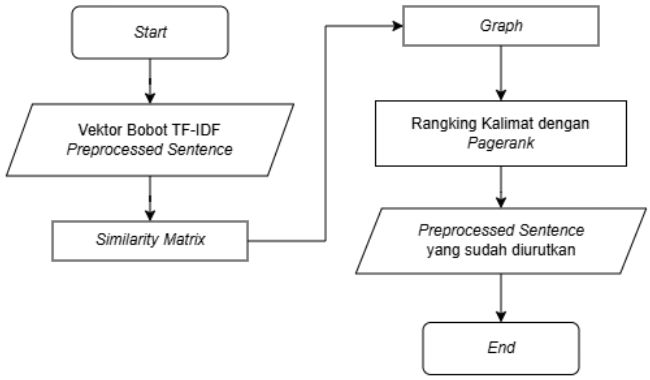
Setelah menyelesaikan tahapan *K-Means Clustering* dan memilih kalimat-kalimat representatif dari setiap *cluster*, hasil tersebut kemudian disusun ulang ke kalimat aslinya untuk membuat hasil ringkasan akhir.

3.6 *Textrank*

TextRank adalah algoritma berbasis graf yang digunakan untuk peringkasan teks otomatis atau ekstraksi informasi dari dokumen. Seperti yang ditunjukkan pada Gambar 3.6, algoritma ini terinspirasi oleh *PageRank*, metode yang awalnya dirancang untuk menentukan peringkat halaman web berdasarkan hubungan antar halaman.

Dalam *flowchart TextRank*, proses dimulai dengan mengambil hasil pembobotan TF-IDF dari tahap sebelumnya sebagai *input*. Setiap kalimat kemudian diwakili sebagai simpul dalam graf, dan kemiripan semantik antar kalimat dihitung menggunakan *cosine similarity* pada representasi vektor TF-IDF nya. Nilai similaritas ini membentuk sisi yang menghubungkan simpul-simpul dalam graf.

Setelah graf terbentuk, algoritma melakukan iterasi perhitungan skor *TextRank* untuk setiap simpul, menggunakan rumus yang diadaptasi dari *PageRank*. Proses iteratif ini berlanjut hingga konvergen atau mencapai jumlah iterasi maksimum yang ditentukan. Pada tahap akhir kalimat-kalimat akan diurutkan berdasarkan skor *TextRank*.



Gambar 3.6 *Flowchart Textrank*

Gambar 3.6 menunjukkan alur proses algoritma *TextRank* yang diterapkan dalam sistem peringkasan dokumen ini. *Flowchart* tersebut menggambarkan tahapan-tahapan utama mulai dari *input* hasil pembobotan TF-IDF hingga perangkingan final yang menghasilkan urutan kalimat optimal untuk ringkasan. Untuk memberikan pemahaman yang lebih mendalam tentang setiap tahapan dalam algoritma *TextRank* ini, berikut akan dijelaskan secara detail proses-proses yang terjadi.

3.6.1 Data *Input*

Tahap awal dalam algoritma *TextRank* adalah mempersiapkan data *input* yang berasal dari hasil pembobotan TF-IDF. Seperti terlihat pada Gambar 3.5, *input* terdiri dari kalimat-kalimat dokumen beserta vektor bobot TF-IDF masing-masing, yang juga digunakan sebagai data *input* untuk metode *K-Means Clustering*. Kalimat-kalimat ini akan diproses lebih lanjut untuk menentukan urutan kepentingannya berdasarkan hubungan semantik antara satu sama lain.

3.6.2 Perhitungan *Cosine Similarity*

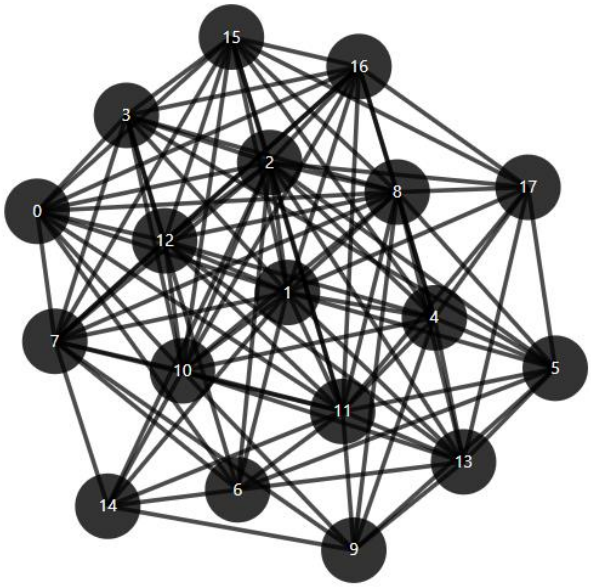
Pada tahap ini, sistem menghitung similaritas antar kalimat menggunakan metrik *cosine similarity*. Perhitungan ini dilakukan untuk setiap pasangan kalimat dengan membandingkan vektor TF-IDF mereka. *Cosine similarit*y dihitung menggunakan Persamaan 2.3. Hasil perhitungan ini menghasilkan nilai antara 0 hingga 1, dimana nilai yang lebih tinggi menunjukkan kemiripan yang lebih besar. Hasil perhitungan *Cosine Similarity* antar kalimat dapat dilihat pada Tabel 3.10

Tabel 3.10 Matriks *Cosine Similarity*

ID	**0**	**1**	**2**	**3**	**…**	**15**	**16**	**17**
0	0.0000	0.0446	0.1887	0.0050	…	0.3486	0.1010	0.0000
1	0.0446	0.0000	0.0491	0.0052	…	0.0490	0.0887	0.0077
2	0.1887	0.0491	0.0000	0.0780	…	0.1110	0.0551	0.0078
3	0.0050	0.0052	0.0780	0.0000	…	0.0055	0.0061	0.0000
…	…	…	…	…	…	…	…	…
15	0.3486	0.0490	0.1110	0.0055	…	0.0000	0.1111	0.0402
16	0.1010	0.0887	0.0551	0.0061	…	0.1111	0.0000	0.0673
17	0.0000	0.0077	0.0078	0.0000	…	0.0402	0.0673	0.0000

3.6.3 Representasi Graf

Pada tahap ini, sistem membangun representasi graf dari kalimat-kalimat yang diproses. Setiap kalimat menjadi sebuah simpul dalam graf, dan nilai *cosine similarity* antar kalimat menjadi bobot sisi yang menghubungkan simpul-simpul tersebut. Misalnya, simpul 0 memiliki hubungan dengan simpul 2 dengan nilai *cosine similarity* 0.1887, yang mengindikasikan adanya kemiripan semantik antara kedua kalimat tersebut. Representasi graf dapat dilihat pada Gambar 3.7



Gambar 3.7 Representasi Graf

Graf yang terbentuk seperti terlihat pada Gambar 3.7 merepresentasikan hubungan semantik antar kalimat dalam dokumen. Setiap simpul mewakili kalimat dalam dokumen, dan garis yang menghubungkan antar simpul menunjukkan adanya kemiripan semantik yang signifikan antar kalimat tersebut. Pada graf ini terlihat kalimat memiliki banyak koneksi yang menandakan kalimat-kalimat tersebut memiliki kemiripan semantik dengan banyak kalimat lain dalam dokumen. Representasi graf ini menjadi dasar untuk perhitungan skor *TextRank* pada tahap selanjutnya.

3.6.4 Perhitungan Skor

Tahap ini melibatkan proses iteratif untuk menghitung skor *TextRank* setiap kalimat. Pertama, sistem menginisialisasi skor awal yang sama untuk setiap kalimat dengan nilai 1/N, dimana N adalah jumlah total kalimat yang diproses. semua

kalimat diberi skor awal 0.05556, yang merupakan hasil dari 1/18 (karena terdapat 18 kalimat yang diproses)

Setelah inisiasi, sistem memulai proses iteratif menggunakan rumus yang diadaptasi dari *PageRank* sebagaimana ditunjukkan pada Persamaan 2.4. Dalam setiap iterasi, skor setiap kalimat diperbarui berdasarkan kontribusi dari kalimat-kalimat yang terhubung dengannya. Proses ini menggunakan *damping factor* 0.85, yang merupakan nilai standar dalam algoritma *Textrank*.

Proses iteratif ini berlanjut hingga mencapai iterasi maksimalnya atau skor-skor telah menunjukkan konvergensi yang cukup signifikan dengan perubahan yang minimal antara iterasi terakhir. Distribusi skor akhir ini mencerminkan tingkat kepentingan relatif setiap kalimat dalam konteks dokumen secara keseluruhan, dengan kalimat-kalimat yang memiliki koneksi semantik lebih kuat ke kalimat-kalimat lain mendapatkan skor yang lebih tinggi.

3.6.5 Perangkingan Kalimat Berdasarkan Skor Akhir

Setelah proses iteratif *TextRank* selesai dan skor final untuk setiap kalimat telah dihitung, sistem mengurutkan kalimat-kalimat berdasarkan skor tersebut dari tertinggi ke terendah. Tabel 3.13 menunjukkan contoh hasil perangkingan kalimat berdasarkan skor *Textrank* akhir yang telah diperoleh.

Tabel 3.11 Contoh Hasil Perangkingan Kalimat

Id	**Kalimat**	**Skor *Textrank***
11	tahun derita sakit lula tak tahu apa sebab mati dr oz indonesia	0.070078

0	jakarta cnn indonesia dokter ryan thamrin kenal lewat acara dokter oz indonesia tinggal dunia jumat dini hari	0.061144
4	tahu orang sehat tahun lalu dengar sakit	0.060390
…	…	…
12	meski dengar beberapa kabar sebut sebab ryan tinggal jatuh kamar mandi	0.048963
3	kondisi buat ryan kampung halaman pekanbaru riau jalan istirahat	0.048121
17	kemudian lanjut didik klinis sehat reproduksi sakit tular seksual mahachulalongkornrajavidyalaya university bangkok thailand	0.046795

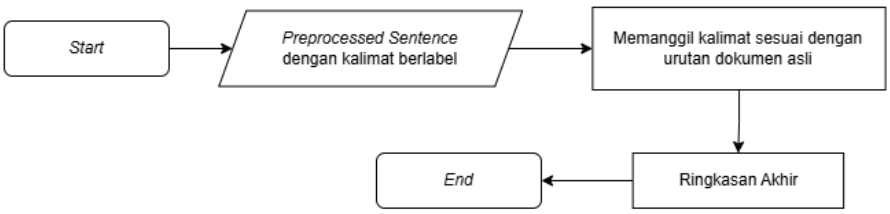
Dari tabel tersebut dapat dilihat bahwa kalimat dengan ID 11 memiliki skor tertinggi yaitu 0.070078, diikuti oleh kalimat ID 0 dengan skor 0.061144. Kalimat-kalimat ini dianggap paling penting atau sentral dalam konteks dokumen berdasarkan hubungan semantiknya dengan kalimat-kalimat lain. Sebagai contoh, kalimat ID 11 yang berbunyi "tahun derita sakit lula tak tahu apa sebab mati dr oz indonesia" mendapat peringkat tertinggi, yang mengindikasikan bahwa kalimat tersebut memiliki koneksi semantik yang kuat dengan banyak kalimat lain dalam dokumen.

Perangkingan ini sangat penting karena menentukan urutan penyajian kalimat dalam ringkasan akhir, memastikan bahwa informasi yang paling penting berdasarkan analisis *TextRank* disajikan lebih dahulu. Dalam implementasi sistem ini, tingkat kompresi yang diterapkan adalah 50%, yang berarti sistem akan memilih

9 kalimat dengan skor tertinggi dari 18 kalimat yang ada sebagai ringkasan akhir. Proses perangkingan ini menghasilkan urutan kalimat yang optimal berdasarkan signifikansi semantiknya, sehingga ringkasan yang dihasilkan terstruktur dengan urutan yang memprioritaskan informasi paling sentral dalam dokumen. Kalimat-kalimat dengan skor *TextRank* tertinggi mencerminkan hubungan paling kuat dengan keseluruhan topik dalam teks asli.

3.7 *Generate Summary*

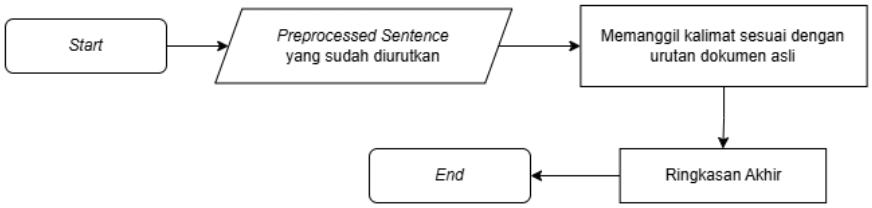
Tahap akhir dalam sistem peringkasan dokumen adalah *Generate Summary*. Untuk metode *K-Means Clustering*, ringkasan dihasilkan dari kalimat-kalimat representatif terpilih dari setiap *cluster*. Sementara untuk metode *TextRank*, ringkasan dibentuk dari kalimat-kalimat dengan skor *TextRank* tertinggi yang disusun berdasarkan urutan kemunculannya dalam dokumen asli. Kedua hasil ringkasan ini kemudian dievaluasi menggunakan metrik ROUGE-1, ROUGE-2, dan ROUGE-L untuk mengukur dan membandingkan kualitas ringkasan yang dihasilkan oleh masing-masing metode.



Gambar 3.8 *Flowchart Generate Summary K-Means Clustering*

Alur proses *generate summary* untuk metode *K-Means Clustering* seperti terlihat pada Gambar 3.8 dimulai dengan input berupa *Preprocessed Sentence* dengan kalimat berlabel. *Input* ini merupakan kalimat-kalimat representatif yang

terpilih dari setiap *cluster*. Meskipun kalimat-kalimat ini telah melalui berbagai tahapan *preprocessing*, sistem tidak menggunakan hasil *preprocessed* untuk ringkasan akhir, namun sistem memanggil kalimat yang sesuai dengan urutan dokumen asli, kemudian sistem membentuk ringkasan akhir.



Gambar 3.9 *Flowchart Generate Summary Textrank*

Sementara itu, alur proses *generate summary* untuk metode *TextRank* seperti terlihat pada Gambar 3.9 dimulai dengan *input* berupa *Preprocessed Sentence* yang sudah diurutkan. *Input* ini berupa kalimat-kalimat dengan skor *TextRank* tertinggi sesuai dengan tingkat kompresi yang ditentukan. Mirip dengan metode *K-Means Clustering*, sistem juga memanggil kalimat yang sesuai dengan urutan dokumen asli, kemudian sistem membentuk ringkasan akhir untuk menghasilkan ringkasan akhir yang koheren. Hasil ringkasan dari metode *K-Means Clustering* dan *TextRank* dapat dilihat pada Tabel 3.12 dan Tabel 3.13

Tabel 3.12 Hasil Ringkasan Sistem *K-Means Clustering*

Compression Rate	**Hasil Ringkasan**
50%	Jakarta, CNN Indonesia -- Dokter Ryan Thamrin, yang terkenal lewat acara Dokter Oz Indonesia, meninggal dunia pada Jumat (4/8) dini hari. "Setahu saya dia orangnya sehat, tapi tahun lalu saya dengar dia sakit. (Karena) sakitnya, ia langsung pulang ke Pekanbaru, jadi kami yang mau jenguk juga susah. Dia juga tak tahu penyakit apa yang diderita Ryan. "Itu saya enggak tahu, belum sempat jenguk dan enggak selamanya bisa dijenguk juga. Enggak tahu berat sekali apa bagaimana, "tutur Ryan. Walau sudah setahun menderita sakit, Lula tak mengetahui apa penyebab pasti kematian Dr Oz Indonesia itu. Kita kan enggak bisa mengambil

	kesimpulan, "kata Lula. Ryan Thamrin terkenal sebagai dokter yang rutin membagikan tips dan informasi kesehatan lewat tayangan Dokter Oz Indonesia.

Tabel 3.12 menampilkan contoh ringkasan yang dihasilkan oleh sistem dengan tingkat kompresi 50% menggunakan metode *K-Means Clustering*. Ringkasan ini merupakan hasil akhir dari rangkaian proses pengelompokan kalimat dan pemilihan kalimat-kalimat representatif dari setiap *cluster*. Dengan menggunakan kalimat-kalimat asli tanpa modifikasi, hasil ringkasan mempertahankan struktur bahasa dan konteks dari dokumen sumber.

Tabel 3.13 Hasil Ringkasan Sistem *TextRank*

Compression Rate	**Hasil Ringkasan**
50%	Jakarta, CNN Indonesia -- Dokter Ryan Thamrin, yang terkenal lewat acara Dokter Oz Indonesia, meninggal dunia pada Jumat (4/8) dini hari. Dokter Lula Kamal yang merupakan selebriti sekaligus rekan kerja Ryan menyebut kawannya itu sudah sakit sejak setahun yang lalu. Lula menuturkan, sakit itu membuat Ryan mesti vakum dari semua kegiatannya, termasuk menjadi pembawa acara Dokter Oz Indonesia. "Setahu saya dia orangnya sehat, tapi tahun lalu saya dengar dia sakit. (Karena) sakitnya, ia langsung pulang ke Pekanbaru, jadi kami yang mau jenguk juga susah. Dia juga tak tahu penyakit apa yang diderita Ryan. "Itu saya enggak tahu, belum sempat jenguk dan enggak selamanya bisa dijenguk juga. Walau sudah setahun menderita sakit, Lula tak mengetahui apa penyebab pasti kematian Dr Oz Indonesia itu. Ryan Thamrin terkenal sebagai dokter yang rutin membagikan tips dan informasi kesehatan lewat tayangan Dokter Oz Indonesia.

Tabel 3.13 menunjukkan contoh ringkasan yang dihasilkan oleh sistem dengan tingkat kompresi 50% menggunakan metode *TextRank*. Ringkasan ini terbentuk dari pemilihan kalimat-kalimat berdasarkan skor *TextRank* tertinggi. Seperti halnya ringkasan dari metode *K-Means Clustering*, hasil ringkasan dari

metode *TextRank* juga menyajikan informasi dari dokumen asli dalam bentuk yang lebih ringkas sesuai dengan tingkat kompresi yang ditentukan.

3.8 Evaluasi Menggunakan Metrik ROUGE

Evaluasi menggunakan metrik ROUGE adalah metode yang banyak digunakan untuk menilai kualitas ringkasan teks otomatis dengan membandingkan ringkasan yang dihasilkan sistem *(generate summary*) dengan ringkasan referensi (*reference summary*). Beberapa metrik utama yang digunakan adalah ROUGE-1, ROUGE-2, dan ROUGE-L. ROUGE-1 menghitung kesesuaian antara *unigram* (kata tunggal) dalam ringkasan sistem dengan ringkasan referensi, sehingga menunjukkan sejauh mana kata-kata penting dari referensi muncul dalam hasil sistem. ROUGE-2 menganalisis kecocokan *bigram* (pasangan kata berurutan), memberikan wawasan tentang koherensi dan struktur kalimat dalam ringkasan. Sementara itu, ROUGE-L mengukur *Longest Common Subsequence* antara ringkasan sistem dan referensi, menangkap fleksibilitas urutan kata dan struktur kalimat yang lebih panjang.

Setiap metrik ROUGE memberikan tiga skor utama yakni *precision*, *recall*, dan *f1-score*. *Precision* mengukur proporsi kata atau sekuens yang relevan dalam ringkasan sistem dibandingkan dengan semua yang dihasilkan oleh sistem, sedangkan *recall* mengukur proporsi elemen dalam ringkasan referensi yang berhasil dicakup oleh sistem. *F1-score* adalah rata-rata harmonis dari *precision* dan *recall*, memberikan keseimbangan antara keduanya. Dalam evaluasi peringkasan teks otomatis, kombinasi ROUGE-1, ROUGE-2, dan ROUGE-L memberikan penilaian yang komprehensif: ROUGE-1 menilai cakupan informasi dasar,

ROUGE-2 menilai kelancaran dan koherensi kalimat, dan ROUGE-L menilai struktur keseluruhan tanpa mengharuskan kecocokan yang berurutan persis. Dengan menggunakan ketiga metrik ini, kualitas ringkasan sistem dapat dinilai secara menyeluruh dari berbagai aspek linguistik. Pada Tabel 3.14 dan Tabel 3.15 akan dipaparkan contoh hasil ROUGE dari masing-masing metode.

Tabel 3.14 Contoh Hasil ROUGE *K-Means Clustering*

Compression Rate	**ROUGE-1**			**ROUGE-2**			**ROUGE-L**		
	Recall	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***
50%	0.415	0.186	0.257	0.077	0.034	0.047	0.226	0.102	0.140

Pada Tabel 3.14 untuk *K-Means Clustering* dengan *compression rate* 50% sebagai rancangan percobaan, terlihat nilai ROUGE-1 mencapai *recall* 0.415, *precision* 0.186, dan *f1-score* 0.257. Ini menunjukkan bahwa metode *K-Means Clustering* mampu mencakup 41.5% kata dari ringkasan referensi, dengan 18.6% kata dalam ringkasan sistem merupakan kata relevan yang juga ada dalam ringkasan referensi. Untuk ROUGE-2, diperoleh nilai *recall* 0.077, *precision* 0.034, dan *f1-score* 0.047, menunjukkan kemampuan metode dalam mempertahankan urutan kata berturut-turut. Sementara untuk ROUGE-L, nilai *recall* mencapai 0.226, *precision* 0.102, dan *f1-score* 0.140. Pada metode ini, kompresi 50% diimplementasikan dengan menentukan jumlah *cluster* K=2.

Tabel 3.15 Contoh Hasil ROUGE *Textrank*

Compression Rate	**ROUGE-1**			**ROUGE-2**			**ROUGE-L**		
	Recall	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***
50%	0.793	0.302	0.438	0.692	0.261	0.379	0.736	0.281	0.406

Sementara, pada Tabel 3.15 hasil evaluasi untuk metode *Textrank* dengan *compression rate* yang sama sebagai rancangan percobaan menunjukkan kinerja yang bervariasi di ketiga metrik ROUGE. Pada ROUGE-1, diperoleh nilai *recall* 0.793, *precision* 0.302, dan *f1-score* 0.438. Hasil ini mengindikasikan bahwa metode *Textrank* berhasil mencakup 79.3% kata dari ringkasan referensi, dengan 30.2% kata dalam ringkasan sistem merupakan kata relevan. Untuk ROUGE-2, nilai *recall* mencapai 0.692, *precision* 0.261, dan *f1-score* 0.379, menunjukkan kemampuan metode dalam mempertahankan pasangan kata berurutan. Adapun untuk ROUGE-L, nilai *recall* 0.736, *precision* 0.281, dan *f1-score* 0.406. Pada metode *Textrank*, kompresi 50% diterapkan dengan memilih kalimat-kalimat yang memiliki skor *Textrank* tertinggi sebanyak 50% dari jumlah total kalimat dalam dokumen.

BAB IV

HASIL DAN PEMBAHASAN

4.1 Skenario Uji Coba

Uji coba dilakukan pada 2500 dokumen yang dipilih dari file *train.03.jsonl* yang merupakan bagian dari dataset IndoSum. Dalam penelitian ini, sistem peringkasan teks diuji dengan menggunakan variasi tingkat kompresi yang berbeda yaitu 30%, 40%, 50%, 60%, dan 70% dari panjang teks asli. Pemilihan variasi tingkat kompresi ini bertujuan untuk menganalisis bagaimana performa sistem dalam menghasilkan ringkasan dengan panjang yang berbeda-beda, serta menemukan titik optimal antara kepadatan informasi dan panjang ringkasan yang dihasilkan. Selain itu, penelitian ini juga melakukan analisis perbandingan antara dua metode peringkasan yang diterapkan, yaitu *K-Means Clustering* dan *TextRank* pada setiap skenario tingkat kompresi untuk mengetahui metode mana yang menghasilkan ringkasan terbaik dalam berbagai kondisi kompresi. Evaluasi performa dilakukan menggunakan tiga metrik ROUGE yang komprehensif yang membandingkan ringkasan sistem dengan ringkasan referensi dari dataset IndoSum. ROUGE-1 mengukur *overlap unigram* (kata tunggal) untuk menilai cakupan konten dasar. ROUGE-2 menganalisis *overlap bigram* (pasangan kata berurutan) untuk mengevaluasi koherensi dan kelancaran kalimat. ROUGE-L mengukur *Longest Common Subsequence* untuk menilai struktur kalimat dan fleksibilitas urutan kata. Ketiga metrik ROUGE ini digunakan untuk menghasilkan nilai *precision* (ketepatan), *recall* (cakupan), dan *f1-score* (keseimbangan antara

precision dan *recall*) pada berbagai tingkat kompresi, memberikan evaluasi menyeluruh terhadap kualitas ringkasan dari berbagai aspek linguistik.

4.2 Pengujian Menggunakan Metode *K-Means Clustering*

Pada pengujian menggunakan metode *K-Means Clustering*, data yang telah melalui tahap pembobotan TF-IDF selanjutnya dikelompokkan ke dalam sejumlah *cluster* berdasarkan kemiripan semantiknya. Setiap *cluster* berisikan kalimat-kalimat dengan karakteristik semantik yang mirip berdasarkan representasi vektor TF-IDF. Sistem kemudian memilih kalimat representatif dari masing-masing *cluster* yang memiliki jarak terdekat dengan *centroid cluster* tersebut. Kalimat-kalimat representatif inilah yang selanjutnya disusun sesuai urutan aslinya dalam dokumen untuk membentuk hasil ringkasan sistem.

4.2.1 Percobaan Skenario 1 dengan Tingkat Kompresi 30%

Pada percobaan skenario ke-1, sistem peringkasan teks diuji dengan menggunakan tingkat kompresi 30%, yang berarti ringkasan yang dihasilkan memiliki panjang 30% dari teks aslinya. Hasil perhitungan nilai ROUGE-1, ROUGE-2, dan ROUGE-L yang masing-masing menghasilkan skor *precision*, *recall*, dan *f1-score* ditampilkan pada Tabel 4.1 berikut

Tabel 4.1 Hasil ROUGE *K-Means Clustering* Skenario 1

Dok.	***Compression Rate* 30%**								
	ROUGE-1			**ROUGE-2**			**ROUGE-L**		
	Recall	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***
1	0.444	0.571	0.500	0.387	0.500	0.436	0.412	0.530	0.464
2	0.507	0.641	0.566	0.393	0.500	0.440	0.447	0.566	0.500
3	0.775	0.405	0.532	0.649	0.336	0.443	0.672	0.351	0.461
4	0.612	0.500	0.550	0.508	0.413	0.455	0.532	0.434	0.478

5	0.730	0.506	0.598	0.490	0.337	0.400	0.596	0.413	0.488
…	…	…	….	….	….	….	….	….	….
2496	0.528	0.493	0.510	0.434	0.405	0.419	0.500	0.466	0.482
2497	0.507	0.761	0.609	0.419	0.634	0.504	0.444	0.666	0.533
2498	0.571	0.628	0.598	0.565	0.513	0.537	0.600	0.545	0.571
2499	0.569	0.521	0.544	0.500	0.457	0.477	0.523	0.478	0.500
2500	0.700	0.560	0.622	0.644	0.513	0.571	0.650	0.520	0.577

Contoh dari hasil ringkasan manual dan sistem pada artikel ke-2, dapat dilihat pada Gambar 4.1, dan Gambar 4.2

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1. Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung.

Gambar 4.1 Ringkasan Manual Artikel-2

BANDUNG, JUARA.net - Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Sebab, di sini lebih besar (kapasitasnya)," kata manajer yang akrab diasapa Pak Haji ini. "Insya Allah itu juga akan kami tertibkan masalah perparkiran.

Gambar 4.2 Ringkasan Sistem *Compression Rate* 30% Artikel-2

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017).

Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1.

Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung selain Stadion Si Jalak Harupat, Kabupaten Bandung.

"Kalau melihat kondisi sekarang, ya lebih baik di GBLA saja. Sebab, di sini lebih besar (kapasitasnya)," kata manajer yang akrab diasapa Pak Haji ini.

Umuh menambahkan, untuk pertandingan perdana, pihaknya saat ini terus mempersiapkan segala sesuatunya terutama masalah keamanan.

Apalagi, pada pertandingan menghadapi tim berjuluk Singo Edan ini, rencanannya Presiden Joko Widodo juga akan hadir langsung di Stadion GBLA.

"Pak Kapolda pun akan turun tangan mengamankan pertandingan tersebut. Selain itu, kalau memang nanti Pak Presiden datang, pasti dari kepresidenan juga akan ikut membantu mengamankan jalannya pertandingan," ucapnya.

Selain pengamanan di dalam Stadion, Umuh juga akan memperketat keamanan di sekitar Stadion dan akan mengambil tindakan tegas jika ada pelanggaran. Pasalnya, pada beberapa laga terakhir, banyak aduan mengenai pungutan liar kepada bobotoh.

"Insya Allah itu juga akan kami tertibkan masalah perparkiran. Jadi, nanti tidak akan ada lagi parkir hingga Rp 20 ribu sampai Rp 50 ribu, karena itu pemerasan," tegas Umuh.

Gambar 4.3 Teks Asli Artikel-2

4.2.2 Percobaan Skenario 2 dengan Tingkat Kompresi 40%

Pada percobaan skenario ke-2, sistem peringkasan teks diuji dengan menggunakan tingkat kompresi 40%, yang berarti ringkasan yang dihasilkan memiliki panjang 40% dari teks aslinya. Hasil perhitungan nilai ROUGE-1, ROUGE-2, dan ROUGE-L yang masing-masing menghasilkan skor *precision*, *recall*, dan *f1-score* ditampilkan pada Tabel 4.2 berikut

Tabel 4.2 Hasil ROUGE *K-Means Clustering* Skenario 2

Dok.	***Compression Rate* 40%**								
	ROUGE-1			**ROUGE-2**			**ROUGE-L**		
	Recall	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***
1	0.537	0.500	0.519	0.419	0.388	0.403	0.476	0.441	0.458
2	0.965	0.423	0.473	0.393	0.309	0.346	0.447	0.352	0.394

3	0.758	0.414	0.580	0.877	0.373	0.523	0.862	0.370	0.518
4	0.884	0.398	0.522	0.622	0.324	0.426	0.709	0.372	0.488
5	0.650	0.410	0.560	0.725	0.333	0.456	0.730	0.339	0.463
…	…	…	….	….	….	….	….	….	….
2496	0.700	0.526	0.601	0.652	0.489	0.559	0.671	0.505	0.576
2497	0.634	0.519	0.571	0.451	0.368	0.405	0.444	0.363	0.4
2498	0.628	0.415	0.500	0.565	0.371	0.448	0.600	0.396	0.477
2499	0.600	0.433	0.503	0.500	0.359	0.418	0.523	0.337	0.438
2500	0.777	0.400	0.525	0.644	0.333	0.439	0.650	0.339	0.445

Contoh dari hasil ringkasan manual dan sistem pada artikel ke-2, dapat dilihat pada Gambar 4.4, dan Gambar 4.5

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1. Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung.

Gambar 4.4 Ringkasan Manual Artikel-2

BANDUNG, JUARA.net - Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Sebab, di sini lebih besar (kapasitasnya), " kata manajer yang akrab diasapa Pak Haji ini. Pasalnya, pada beberapa laga terakhir, banyak aduan mengenai pungutan liar kepada bobotoh. "Insya Allah itu juga akan kami tertibkan masalah perparkiran. Jadi, nanti tidak akan ada lagi parkir hingga Rp 20 ribu sampai Rp 50 ribu, karena itu pemerasan," tegas Umuh.

Gambar 4.5 Ringkasan Sistem *Compression Rate* 40% Artikel-2

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017).

Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1.

Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung selain Stadion Si Jalak Harupat, Kabupaten Bandung.

"Kalau melihat kondisi sekarang, ya lebih baik di GBLA saja. Sebab, di sini lebih besar (kapasitasnya)," kata manajer yang akrab diasapa Pak Haji ini.

Umuh menambahkan, untuk pertandingan perdana, pihaknya saat ini terus mempersiapkan segala sesuatunya terutama masalah keamanan.

Apalagi, pada pertandingan menghadapi tim berjuluk Singo Edan ini, rencanannya Presiden Joko Widodo juga akan hadir langsung di Stadion GBLA.

"Pak Kapolda pun akan turun tangan mengamankan pertandingan tersebut. Selain itu, kalau memang nanti Pak Presiden datang, pasti dari kepresidenan juga akan ikut membantu mengamankan jalannya pertandingan," ucapnya.

Selain pengamanan di dalam Stadion, Umuh juga akan memperketat keamanan di sekitar Stadion dan akan mengambil tindakan tegas jika ada pelanggaran. Pasalnya, pada beberapa laga terakhir, banyak aduan mengenai pungutan liar kepada bobotoh.

"Insya Allah itu juga akan kami tertibkan masalah perparkiran. Jadi, nanti tidak akan ada lagi parkir hingga Rp 20 ribu sampai Rp 50 ribu, karena itu pemerasan," tegas Umuh.

Gambar 4.6 Teks asli artikel-2

4.2.3 Percobaan Skenario 3 dengan Tingkat Kompresi 50%

Pada percobaan skenario ke-3, sistem peringkasan teks diuji dengan menggunakan tingkat kompresi 50%, yang berarti ringkasan yang dihasilkan memiliki panjang 50% dari teks aslinya. Hasil perhitungan nilai ROUGE-1, ROUGE-2, dan ROUGE-L yang masing-masing menghasilkan skor *precision*, *recall*, dan *f1-score* ditampilkan pada Tabel 4.3 berikut

Tabel 4.3 Hasil ROUGE *K-Means Clustering* Skenario 3

Dok.	***Compression Rate* 50%**								
	ROUGE-1			**ROUGE-2**			**ROUGE-L**		
	Recall	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***
1	0.777	0.550	0.644	0.693	0.488	0.573	0.698	0.494	0.578
2	0.761	0.472	0.582	0.621	0.383	0.473	0.641	0.398	0.491

3	0.965	0.343	0.506	0.877	0.308	0.456	0.862	0.306	0.452
4	0.758	0.297	0.427	0.622	0.242	0.348	0.709	0.278	0.400
5	0.884	0.368	0.519	0.705	0.290	0.411	0.730	0.304	0.429
…	…	…	….	….	….	….	….	….	….
2496	0.728	0.443	0.551	0.666	0.502	0.502	0.700	0.426	0.529
2497	0.666	0.403	0.502	0.451	0.339	0.339	0.476	0.288	0.359
2498	0.628	0.352	0.451	0.565	0.404	0.404	0.600	0.336	0.430
2499	0.830	0.482	0.610	0.781	0.571	0.571	0.815	0.473	0.598
2500	1	0.431	0.603	1	0.598	0.598	1	0.431	0.603

Contoh dari hasil ringkasan manual dan sistem pada artikel ke-2, dapat dilihat pada Gambar 4.7, dan Gambar 4.8

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1. Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung.

Gambar 4.7 Ringkasan Manual Artikel-2

BANDUNG, JUARA.net - Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung selain Stadion Si Jalak Harupat, Kabupaten Bandung. Sebab, di sini lebih besar (kapasitasnya)," kata manajer yang akrab diasapa Pak Haji ini. Pasalnya, pada beberapa laga terakhir, banyak aduan mengenai pungutan liar kepada bobotoh. "Insya Allah itu juga akan kami tertibkan masalah perparkiran. Jadi, nanti tidak akan ada lagi parkir hingga Rp 20 ribu sampai Rp 50 ribu, karena itu pemerasan," tegas Umuh.

Gambar 4.8 Ringkasan Sistem *Compression Rate* 50% Artikel-2

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017).

Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1.

Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung selain Stadion Si Jalak Harupat, Kabupaten Bandung.

"Kalau melihat kondisi sekarang, ya lebih baik di GBLA saja. Sebab, di sini lebih besar (kapasitasnya)," kata manajer yang akrab diasapa Pak Haji ini.

Umuh menambahkan, untuk pertandingan perdana, pihaknya saat ini terus mempersiapkan segala sesuatunya terutama masalah keamanan.

Apalagi, pada pertandingan menghadapi tim berjuluk Singo Edan ini, rencanannya Presiden Joko Widodo juga akan hadir langsung di Stadion GBLA.

"Pak Kapolda pun akan turun tangan mengamankan pertandingan tersebut. Selain itu, kalau memang nanti Pak Presiden datang, pasti dari kepresidenan juga akan ikut membantu mengamankan jalannya pertandingan," ucapnya.

Selain pengamanan di dalam Stadion, Umuh juga akan memperketat keamanan di sekitar Stadion dan akan mengambil tindakan tegas jika ada pelanggaran. Pasalnya, pada beberapa laga terakhir, banyak aduan mengenai pungutan liar kepada bobotoh.

"Insya Allah itu juga akan kami tertibkan masalah perparkiran. Jadi, nanti tidak akan ada lagi parkir hingga Rp 20 ribu sampai Rp 50 ribu, karena itu pemerasan," tegas Umuh.

Gambar 4.9 Teks asli artikel-2

4.2.4 Percobaan Skenario 4 dengan Tingkat Kompresi 60%

Pada percobaan skenario ke-4, sistem peringkasan teks diuji dengan menggunakan tingkat kompresi 60%, yang berarti ringkasan yang dihasilkan memiliki panjang 60% dari teks aslinya. Hasil perhitungan nilai ROUGE-1, ROUGE-2, dan ROUGE-L yang masing-masing menghasilkan skor *precision*, *recall*, dan *f1-score* ditampilkan pada Tabel 4.4 berikut

Tabel 4.4 Hasil ROUGE *K-Means Clustering* Skenario 4

Dok.	***Compression Rate* 60%**								
	ROUGE-1			**ROUGE-2**			**ROUGE-L**		
	Recall	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***
1	0.793	0.510	0.621	0.693	0.443	0.540	0.448	0.698	0.546
2	0.776	0.406	0.533	0.621	0.332	0.428	0.335	0.641	0.441

3	1	0.320	0.485	0.964	0.305	0.464	0.320	1	0.485
4	0.919	0.308	0.461	0.852	0.282	0.424	0.291	0.870	0.437
5	0.980	0.320	0.483	0.960	0.310	0.468	0.320	0.980	0.483
…	…	…	….	….	….	….	….	….	….
2496	0.757	0.381	0.507	0.666	0.333	0.444	0.728	0.366	0.488
2497	0.793	0.431	0.558	0.661	0.356	0.463	0.666	0.362	0.469
2498	0.657	0.312	0.423	0.565	0.267	0.362	0.6	0.285	0.387
2499	0.830	0.400	0.540	0.781	0.373	0.505	0.815	0.392	0.530
2500	1	0.363	0.533	1	0.359	0.529	1	0.363	0.533

Contoh dari hasil ringkasan manual dan sistem pada artikel ke-2, dapat dilihat pada Gambar 4.10, dan Gambar 4.11

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1. Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung.

Gambar 4.10 Ringkasan Manual Artikel-2

BANDUNG, JUARA.net - Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung selain Stadion Si Jalak Harupat, Kabupaten Bandung. Sebab, di sini lebih besar (kapasitasnya), " kata manajer yang akrab diasapa Pak Haji ini. Apalagi, pada pertandingan menghadapi tim berjuluk Singo Edan ini, rencanannya Presiden Joko Widodo juga akan hadir langsung di Stadion GBLA. Pasalnya, pada beberapa laga terakhir, banyak aduan mengenai pungutan liar kepada bobotoh. "Insya Allah itu juga akan kami tertibkan masalah perparkiran. Jadi, nanti tidak akan ada lagi parkir hingga Rp 20 ribu sampai Rp 50 ribu, karena itu pemerasan," tegas Umuh.

Gambar 4.11 Ringkasan Sistem *Compression Rate* 60% Artikel-2

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017).

Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1.

Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung selain Stadion Si Jalak Harupat, Kabupaten Bandung.

"Kalau melihat kondisi sekarang, ya lebih baik di GBLA saja. Sebab, di sini lebih besar (kapasitasnya)," kata manajer yang akrab diasapa Pak Haji ini.

Umuh menambahkan, untuk pertandingan perdana, pihaknya saat ini terus mempersiapkan segala sesuatunya terutama masalah keamanan.

Apalagi, pada pertandingan menghadapi tim berjuluk Singo Edan ini, rencanannya Presiden Joko Widodo juga akan hadir langsung di Stadion GBLA.

"Pak Kapolda pun akan turun tangan mengamankan pertandingan tersebut. Selain itu, kalau memang nanti Pak Presiden datang, pasti dari kepresidenan juga akan ikut membantu mengamankan jalannya pertandingan," ucapnya.

Selain pengamanan di dalam Stadion, Umuh juga akan memperketat keamanan di sekitar Stadion dan akan mengambil tindakan tegas jika ada pelanggaran. Pasalnya, pada beberapa laga terakhir, banyak aduan mengenai pungutan liar kepada bobotoh.

"Insya Allah itu juga akan kami tertibkan masalah perparkiran. Jadi, nanti tidak akan ada lagi parkir hingga Rp 20 ribu sampai Rp 50 ribu, karena itu pemerasan," tegas Umuh.

Gambar 4.12 Teks asli artikel-2

4.2.5 Percobaan Skenario 5 dengan Tingkat Kompresi 70%

Pada percobaan skenario ke-5, sistem peringkasan teks diuji dengan menggunakan tingkat kompresi 70%, yang berarti ringkasan yang dihasilkan memiliki panjang 70% dari teks aslinya. Hasil perhitungan nilai ROUGE-1, ROUGE-2, dan ROUGE-L yang masing-masing menghasilkan skor *precision*, *recall*, dan *f1-score* ditampilkan pada Tabel 4.5 berikut

Tabel 4.5 Hasil ROUGE *K-Means Clustering* Skenario 5

Dok.	***Compression Rate* 70%**								
	ROUGE-1			**ROUGE-2**			**ROUGE-L**		
	Recall	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***
1	1	0.486	0.663	0.967	0.476	0.638	0.730	0.362	0.484
2	1	0.391	0.563	1	0.338	0.559	1	0.391	0.563

3	1	0.288	0.447	0.964	0.275	0.428	1	0.288	0.447
4	0.919	0.276	0.425	0.852	0.253	0.390	0.870	0.262	0.402
5	0.980	0.309	0.470	0.960	0.298	0.455	0.980	0.309	0.470
…	…	…	….	….	….	….	….	….	….
2496	0.757	0.355	0.484	0.666	0.310	0.423	0.728	0.342	0.465
2497	0.904	0.341	0.495	0.741	0.277	0.403	0.825	0.311	0.452
2498	0.871	0.369	0.519	0.811	0.341	0.480	0.842	0.357	0.502
2499	1	0.427	0.599	0.968	0.410	0.576	1	0.427	0.599
2500	1	0.312	0.476	1	0.308	0.472	1	0.312	0.476

Contoh dari hasil ringkasan manual dan sistem pada artikel ke-2, dapat dilihat pada Gambar 4.13, dan Gambar 4.14

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1. Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung.

Gambar 4.13 Ringkasan Manual Artikel-2

BANDUNG, JUARA.net - Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1. Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung selain Stadion Si Jalak Harupat, Kabupaten Bandung. Sebab, di sini lebih besar (kapasitasnya)," kata manajer yang akrab diasapa Pak Haji ini. Apalagi, pada pertandingan menghadapi tim berjuluk Singo Edan ini, rencanannya Presiden Joko Widodo juga akan hadir langsung di Stadion GBLA. Selain itu, kalau memang nanti Pak Presiden datang, pasti dari kepresidenan juga akan ikut membantu mengamankan jalannya pertandingan," ucapnya. Pasalnya, pada beberapa laga terakhir, banyak aduan mengenai pungutan liar kepada bobotoh. "Insya Allah itu juga akan kami tertibkan masalah perparkiran. Jadi, nanti tidak akan ada lagi parkir hingga Rp 20 ribu sampai Rp 50 ribu, karena itu pemerasan," tegas Umuh.

Gambar 4.14 Ringkasan Sistem *Compression Rate* 70% Artikel-2

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017).

Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1.

Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung selain Stadion Si Jalak Harupat, Kabupaten Bandung.

"Kalau melihat kondisi sekarang, ya lebih baik di GBLA saja. Sebab, di sini lebih besar (kapasitasnya)," kata manajer yang akrab diasapa Pak Haji ini.

Umuh menambahkan, untuk pertandingan perdana, pihaknya saat ini terus mempersiapkan segala sesuatunya terutama masalah keamanan.

Apalagi, pada pertandingan menghadapi tim berjuluk Singo Edan ini, rencanannya Presiden Joko Widodo juga akan hadir langsung di Stadion GBLA.

"Pak Kapolda pun akan turun tangan mengamankan pertandingan tersebut. Selain itu, kalau memang nanti Pak Presiden datang, pasti dari kepresidenan juga akan ikut membantu mengamankan jalannya pertandingan," ucapnya.

Selain pengamanan di dalam Stadion, Umuh juga akan memperketat keamanan di sekitar Stadion dan akan mengambil tindakan tegas jika ada pelanggaran. Pasalnya, pada beberapa laga terakhir, banyak aduan mengenai pungutan liar kepada bobotoh.

"Insya Allah itu juga akan kami tertibkan masalah perparkiran. Jadi, nanti tidak akan ada lagi parkir hingga Rp 20 ribu sampai Rp 50 ribu, karena itu pemerasan," tegas Umuh.

Gambar 4.15 Teks Asli Artikel-2

4.3 Pengujian Menggunakan Metode *Textrank*

Pada pengujian menggunakan metode *TextRank*, data yang telah melalui tahap pembobotan TF-IDF selanjutnya dikonversi menjadi graf kalimat dimana setiap kalimat diwakili sebagai simpul dalam graf. Hubungan antar kalimat dihitung menggunakan *cosine similarity* pada representasi vektor TF-IDF, yang membentuk sisi dengan bobot tertentu pada graf. Algoritma *TextRank* kemudian diterapkan untuk memberikan skor pada setiap kalimat berdasarkan pentingnya dalam graf tersebut. Kalimat-kalimat dengan skor tertinggi dipilih sesuai dengan tingkat kompresi yang ditentukan, dan akhirnya disusun sesuai urutan aslinya dalam dokumen untuk membentuk hasil ringkasan sistem.

4.3.1 Percobaan Skenario 1 dengan Tingkat Kompresi 30%

Pada percobaan skenario ke-1, sistem peringkasan teks diuji dengan menggunakan tingkat kompresi 30%, yang berarti ringkasan yang dihasilkan memiliki panjang 30% dari teks aslinya. Hasil perhitungan nilai ROUGE-1, ROUGE-2, dan ROUGE-L yang masing-masing menghasilkan skor *precision*, *recall*, dan *f1-score* ditampilkan pada Tabel 4.6 berikut.

Tabel 4.6 Hasil ROUGE *Textrank* Skenario 1

Dok.	***Compression Rate 30%***								
	ROUGE-1			**ROUGE-2**			**ROUGE-L**		
	Recall	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***
1	0.761	0.727	0.744	0.661	0.630	0.645	0.428	0.409	0.418
2	0.462	0.516	0.488	0.363	0.406	0.384	0.388	0.433	0.409
3	0.737	0.505	0.600	0.700	0.477	0.567	0.721	0.494	0.586
4	0.965	0.459	0.622	0.877	0.413	0.561	0.862	0.409	0.555
5	0.682	0.614	0.646	0.629	0.565	0.595	0.666	0.600	0.631
…	…	…	….	….	….	….	….	….	….
2496	0.666	0.389	0.654	0.553	0.321	0.576	0.575	0.336	0.424
2497	0.576	0.428	0.491	0.450	0.333	0.406	0.403	0.300	0.344
2498	0.803	0.576	0.491	0.727	0.519	0.383	0.732	0.525	0.611
2499	0.777	0.500	0.671	0.622	0.397	0.606	0.648	0.416	0.507
2500	0.761	0.564	0.608	0.661	0.488	0.485	0.444	0.329	0.378

Contoh dari hasil ringkasan manual dan sistem pada artikel ke-2, dapat dilihat pada Gambar 4.16, dan Gambar 4.17

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1. Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung.

Gambar 4.16 Ringkasan Manual Artikel-2

Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1. Umuh menambahkan, untuk pertandingan perdana, pihaknya saat ini terus mempersiapkan segala sesuatunya terutama masalah keamanan. Selain pengamanan di dalam Stadion, Umuh juga akan memperketat keamanan di sekitar Stadion dan akan mengambil tindakan tegas jika ada pelanggaran.

Gambar 4.17 Ringkasan Sistem *Compression Rate* 30% Artikel-2

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017).

Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1.

Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung selain Stadion Si Jalak Harupat, Kabupaten Bandung.

"Kalau melihat kondisi sekarang, ya lebih baik di GBLA saja. Sebab, di sini lebih besar (kapasitasnya)," kata manajer yang akrab diasapa Pak Haji ini.

Umuh menambahkan, untuk pertandingan perdana, pihaknya saat ini terus mempersiapkan segala sesuatunya terutama masalah keamanan.

Apalagi, pada pertandingan menghadapi tim berjuluk Singo Edan ini, rencanannya Presiden Joko Widodo juga akan hadir langsung di Stadion GBLA.

"Pak Kapolda pun akan turun tangan mengamankan pertandingan tersebut. Selain itu, kalau memang nanti Pak Presiden datang, pasti dari kepresidenan juga akan ikut membantu mengamankan jalannya pertandingan," ucapnya.

Selain pengamanan di dalam Stadion, Umuh juga akan memperketat keamanan di sekitar Stadion dan akan mengambil tindakan tegas jika ada pelanggaran. Pasalnya, pada beberapa laga terakhir, banyak aduan mengenai pungutan liar kepada bobotoh.

"Insya Allah itu juga akan kami tertibkan masalah perparkiran. Jadi, nanti tidak akan ada lagi parkir hingga Rp 20 ribu sampai Rp 50 ribu, karena itu pemerasan," tegas Umuh.

Gambar 4.18 Teks Asli Artikel-2

4.3.2 Percobaan Skenario 2 dengan Tingkat Kompresi 40%

Pada percobaan skenario ke-2, sistem peringkasan teks diuji dengan menggunakan tingkat kompresi 40%, yang berarti ringkasan yang dihasilkan memiliki panjang 40% dari teks aslinya. Hasil perhitungan nilai ROUGE-1,

ROUGE-2, dan ROUGE-L yang masing-masing menghasilkan skor *precision*, *recall*, dan *f1-score* ditampilkan pada Tabel 4.7 berikut.

Tabel 4.7 Hasil ROUGE *TextRank* Skenario 2

Dok.	***Compression Rate* 40%**								
	ROUGE-1			**ROUGE-2**			**ROUGE-L**		
	Recall	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***
1	0.777	0.500	0.648	0.661	0.488	0.561	0.444	0.329	0.378
2	0.761	0.564	0.625	0.757	0.462	0.574	0.791	0.486	0.602
3	0.820	0.504	0.466	0.700	0.320	0.439	0.721	0.333	0.455
4	0.737	0.340	0.568	0.877	0.362	0.512	0.862	0.359	0.507
5	0.965	0.402	0.592	0.629	0.443	0.520	0.682	0.483	0.565
…	…	…	….	….	….	….	….	….	….
2496	0.587	0.560	0.578	0.516	0.500	0.507	0.539	0.523	0.531
2497	1	0.431	0.602	0.984	0.421	0.589	1	0.431	0.602
2498	0.750	0.364	0.490	0.666	0.320	0.433	0.596	0.289	0.389
2499	0.803	0.483	0.604	0.727	0.434	0.544	0.732	0.440	0.550
2500	0.925	0.423	0.581	0.867	0.393	0.541	0.907	0.415	0.569

Contoh dari hasil ringkasan manual dan sistem pada artikel ke-2, dapat dilihat pada Gambar 4.19, dan Gambar 4.20

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1. Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung.

Gambar 4.19 Ringkasan Manual Artikel-2

BANDUNG, JUARA.net - Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1. Umuh menambahkan, untuk pertandingan perdana, pihaknya saat ini terus mempersiapkan segala sesuatunya terutama masalah keamanan. Selain itu, kalau memang nanti Pak Presiden datang, pasti dari kepresidenan juga akan ikut membantu mengamankan jalannya pertandingan, " ucapnya. Selain pengamanan di dalam Stadion, Umuh juga akan memperketat keamanan di sekitar Stadion dan akan mengambil tindakan tegas jika ada pelanggaran.

Gambar 4.20 Ringkasan Sistem *Compression Rate* 40% Artikel-2

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017).

Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1.

Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung selain Stadion Si Jalak Harupat, Kabupaten Bandung.

"Kalau melihat kondisi sekarang, ya lebih baik di GBLA saja. Sebab, di sini lebih besar (kapasitasnya)," kata manajer yang akrab diasapa Pak Haji ini.

Umuh menambahkan, untuk pertandingan perdana, pihaknya saat ini terus mempersiapkan segala sesuatunya terutama masalah keamanan.

Apalagi, pada pertandingan menghadapi tim berjuluk Singo Edan ini, rencanannya Presiden Joko Widodo juga akan hadir langsung di Stadion GBLA.

"Pak Kapolda pun akan turun tangan mengamankan pertandingan tersebut. Selain itu, kalau memang nanti Pak Presiden datang, pasti dari kepresidenan juga akan ikut membantu mengamankan jalannya pertandingan," ucapnya.

Selain pengamanan di dalam Stadion, Umuh juga akan memperketat keamanan di sekitar Stadion dan akan mengambil tindakan tegas jika ada pelanggaran. Pasalnya, pada beberapa laga terakhir, banyak aduan mengenai pungutan liar kepada bobotoh.

"Insya Allah itu juga akan kami tertibkan masalah perparkiran. Jadi, nanti tidak akan ada lagi parkir hingga Rp 20 ribu sampai Rp 50 ribu, karena itu pemerasan," tegas Umuh.

Gambar 4.21 Teks Asli Artikel-2

4.3.3 Percobaan Skenario 3 dengan Tingkat Kompresi 50%

Pada percobaan skenario ke-3, sistem peringkasan teks diuji dengan menggunakan tingkat kompresi 50%, yang berarti ringkasan yang dihasilkan memiliki panjang 50% dari teks aslinya. Hasil perhitungan nilai ROUGE-1, ROUGE-2, dan ROUGE-L yang masing-masing menghasilkan skor *precision*, *recall*, dan *f1-score* ditampilkan pada Tabel 4.8 berikut.

Tabel 4.8 Hasil ROUGE *TextRank* Skenario 3

Dok.	***Compression Rate* 50%**		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**

	Recall	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***
1	0.777	0.453	0.573	0.661	0.383	0.485	0.476	0.277	0.350
2	0.820	0.426	0.561	0.757	0.390	0.515	0.791	0.410	0.540
3	0.901	0.384	0.539	0.883	0.373	0.524	0.901	0.384	0.539
4	0.965	0.375	0.541	0.877	0.337	0.487	0.862	0.335	0.483
5	0.730	0.403	0.519	0.629	0.345	0.445	0.682	0.377	0.485
…	…	…	….	….	….	….	….	….	….
2496	0.603	0.475	0.531	0.516	0.405	0.453	0.539	0.425	0.475
2497	1	0.362	0.532	0.984	0.353	0.520	1	0.362	0.532
2498	1	0.363	0.533	1	0.359	0.528	1	0.363	0.533
2499	0.803	0.364	0.520	0.727	0.344	0.467	0.732	0.350	0.473
2500	0.925	0.362	0.520	0.867	0.335	0.484	0.907	0.355	0.514

Contoh dari hasil ringkasan manual dan sistem pada artikel ke-4, dapat dilihat pada Gambar 4.22, dan Gambar 4.23

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1. Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung.

Gambar 4.22 Ringkasan Manual Artikel-2

BANDUNG, JUARA.net - Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1. Umuh menambahkan, untuk pertandingan perdana, pihaknya saat ini terus mempersiapkan segala sesuatunya terutama masalah keamanan. Apalagi, pada pertandingan menghadapi tim berjuluk Singo Edan ini, rencanannya Presiden Joko Widodo juga akan hadir langsung di Stadion GBLA. Selain itu, kalau memang nanti Pak Presiden datang, pasti dari kepresidenan juga akan ikut membantu mengamankan jalannya pertandingan, " ucapnya. Selain pengamanan di dalam Stadion, Umuh juga akan memperketat keamanan di sekitar Stadion dan akan mengambil tindakan tegas jika ada pelanggaran.

Gambar 4.23 Ringkasan Sistem *Compression Rate* 50% Artikel-2

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017).

Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1.

Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung selain Stadion Si Jalak Harupat, Kabupaten Bandung.

"Kalau melihat kondisi sekarang, ya lebih baik di GBLA saja. Sebab, di sini lebih besar (kapasitasnya)," kata manajer yang akrab diasapa Pak Haji ini.

Umuh menambahkan, untuk pertandingan perdana, pihaknya saat ini terus mempersiapkan segala sesuatunya terutama masalah keamanan.

Apalagi, pada pertandingan menghadapi tim berjuluk Singo Edan ini, rencanannya Presiden Joko Widodo juga akan hadir langsung di Stadion GBLA.

"Pak Kapolda pun akan turun tangan mengamankan pertandingan tersebut. Selain itu, kalau memang nanti Pak Presiden datang, pasti dari kepresidenan juga akan ikut membantu mengamankan jalannya pertandingan," ucapnya.

Selain pengamanan di dalam Stadion, Umuh juga akan memperketat keamanan di sekitar Stadion dan akan mengambil tindakan tegas jika ada pelanggaran. Pasalnya, pada beberapa laga terakhir, banyak aduan mengenai pungutan liar kepada bobotoh.

"Insya Allah itu juga akan kami tertibkan masalah perparkiran. Jadi, nanti tidak akan ada lagi parkir hingga Rp 20 ribu sampai Rp 50 ribu, karena itu pemerasan," tegas Umuh.

Gambar 4.24 Teks Asli Artikel-2

4.3.4 Percobaan Skenario 4 dengan Tingkat Kompresi 60%

Pada percobaan skenario ke-4, sistem peringkasan teks diuji dengan menggunakan tingkat kompresi 60%, yang berarti ringkasan yang dihasilkan memiliki panjang 60% dari teks aslinya. Hasil perhitungan nilai ROUGE-1, ROUGE-2, dan ROUGE-L yang masing-masing menghasilkan skor *precision*, *recall*, dan *f1-score* ditampilkan pada Tabel 4.9 berikut.

Tabel 4.9 Hasil ROUGE *TextRank* Skenario 4

Dok.	***Compression Rate* 60%**								
	ROUGE-1			**ROUGE-2**			**ROUGE-L**		
	Recall	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***
1	1	0.473	0.642	0.967	0.454	0.618	0.730	0.345	0.469
2	1	0.440	0.611	1	0.437	0.608	1	0.440	0.611

3	0.901	0.329	0.482	0.883	0.319	0.469	0.901	0.329	0.482
4	0.965	0.321	0.482	0.877	0.289	0.434	0.862	0.287	0.431
5	0.825	0.422	0.559	0.774	0.393	0.521	0.793	0.406	0.537
…	…	…	….	….	….	….	….	….	….
2496	1	0.577	0.732	0.967	0.555	0.705	1	0.557	0.732
2497	1	0.318	0.483	0.984	0.310	0.472	1	0.318	0.483
2498	1	0.337	0.504	1	0.333	0.500	1	0.337	0.504
2499	1	0.366	0.535	1	0.361	0.531	1	0.366	0.535
2500	0.925	0.314	0.469	0.867	0.291	0.436	0.907	0.308	0.460

Contoh dari hasil ringkasan manual dan sistem pada artikel ke-4, dapat dilihat pada Gambar 4.25, dan Gambar 4.26

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1. Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung.

Gambar 4.25 Ringkasan Manual Artikel-2

BANDUNG, JUARA.net - Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1. Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung selain Stadion Si Jalak Harupat, Kabupaten Bandung. Umuh menambahkan, untuk pertandingan perdana, pihaknya saat ini terus mempersiapkan segala sesuatunya terutama masalah keamanan. Apalagi, pada pertandingan menghadapi tim berjuluk Singo Edan ini, rencanannya Presiden Joko Widodo juga akan hadir langsung di Stadion GBLA. Selain itu, kalau memang nanti Pak Presiden datang, pasti dari kepresidenan juga akan ikut membantu mengamankan jalannya pertandingan, " ucapnya. Selain pengamanan di dalam Stadion, Umuh juga akan memperketat keamanan di sekitar Stadion dan akan mengambil tindakan tegas jika ada pelanggaran.

Gambar 4.26 Ringkasan Sistem *Compression Rate* 60% Artikel-2

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017).

Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1.

Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung selain Stadion Si Jalak Harupat, Kabupaten Bandung.

"Kalau melihat kondisi sekarang, ya lebih baik di GBLA saja. Sebab, di sini lebih besar (kapasitasnya)," kata manajer yang akrab diasapa Pak Haji ini.

Umuh menambahkan, untuk pertandingan perdana, pihaknya saat ini terus mempersiapkan segala sesuatunya terutama masalah keamanan.

Apalagi, pada pertandingan menghadapi tim berjuluk Singo Edan ini, rencanannya Presiden Joko Widodo juga akan hadir langsung di Stadion GBLA.

"Pak Kapolda pun akan turun tangan mengamankan pertandingan tersebut. Selain itu, kalau memang nanti Pak Presiden datang, pasti dari kepresidenan juga akan ikut membantu mengamankan jalannya pertandingan," ucapnya.

Selain pengamanan di dalam Stadion, Umuh juga akan memperketat keamanan di sekitar Stadion dan akan mengambil tindakan tegas jika ada pelanggaran. Pasalnya, pada beberapa laga terakhir, banyak aduan mengenai pungutan liar kepada bobotoh.

"Insya Allah itu juga akan kami tertibkan masalah perparkiran. Jadi, nanti tidak akan ada lagi parkir hingga Rp 20 ribu sampai Rp 50 ribu, karena itu pemerasan," tegas Umuh.

Gambar 4.27 Teks Asli Artikel-2

4.3.5 Percobaan Skenario 5 dengan Tingkat Kompresi 70%

Pada percobaan skenario ke-5, sistem peringkasan teks diuji dengan menggunakan tingkat kompresi 70%, yang berarti ringkasan yang dihasilkan memiliki panjang 70% dari teks aslinya. Hasil perhitungan nilai ROUGE-1, ROUGE-2, dan ROUGE-L yang masing-masing menghasilkan skor *precision*, *recall*, dan *f1-score* ditampilkan pada Tabel 4.10 berikut.

Tabel 4.10 Hasil ROUGE *TextRank* Skenario 5

Dok.	***Compression Rate* 70%**								
	ROUGE-1			**ROUGE-2**			**ROUGE-L**		
	Recall	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***	***Recall***	***Precision***	***F1-Score***
1	1	0.443	0.614	0.967	0.425	0.591	0.730	0.323	0.448
2	1	0.382	0.553	1	0.379	0.550	1	0.382	0.553

3	0.901	0.277	0.424	0.883	0.269	0.412	0.901	0.277	0.424
4	1	0.300	0.462	0.964	0.286	0.441	1	0.300	0.462
5	0.841	0.353	0.497	0.774	0.322	0.454	0.793	0.333	0.491
…	…	…	….	….	….	….	….	….	….
2496	1	0.470	0.639	0.967	0.451	0.615	1	0.470	0.639
2497	1	0.275	0.431	0.984	0.267	0.421	1	0.275	0.431
2498	1	0.285	0.444	1	0.281	0.439	1	0.285	0.444
2499	1	0.335	0.502	1	0.331	0.497	1	0.335	0.502
2500	0.925	0.257	0.403	0.867	0.238	0.373	0.907	0.252	0.395

Contoh dari hasil ringkasan manual dan sistem pada artikel ke-4, dapat dilihat pada Gambar 4.28, dan Gambar 4.29

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1. Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung.

Gambar 4.28 Ringkasan Manual Artikel-2

BANDUNG, JUARA.net - Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017). Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1. Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung selain Stadion Si Jalak Harupat, Kabupaten Bandung. Sebab, di sini lebih besar (kapasitasnya)," kata manajer yang akrab diasapa Pak Haji ini. Umuh menambahkan, untuk pertandingan perdana, pihaknya saat ini terus mempersiapkan segala sesuatunya terutama masalah keamanan. Apalagi, pada pertandingan menghadapi tim berjuluk Singo Edan ini, rencanannya Presiden Joko Widodo juga akan hadir langsung di Stadion GBLA. "Pak Kapolda pun akan turun tangan mengamankan pertandingan tersebut. Selain itu, kalau memang nanti Pak Presiden datang, pasti dari kepresidenan juga akan ikut membantu mengamankan jalannya pertandingan," ucapnya. Selain pengamanan di dalam Stadion, Umuh juga akan memperketat keamanan di sekitar Stadion dan akan mengambil tindakan tegas jika ada pelanggaran.

Gambar 4.29 Ringkasan Sistem *Compression Rate* 70% Artikel-2

Persib Bandung dipastikan bakal menggunakan Stadion Gelora Bandung Lautan Api (GBLA), Kota Bandung saat menjamu Arema FC pada pertandingan perdana kompetisi Liga 1, Sabtu (15/4/2017).

Selain itu, menurut manajer Persib Umuh Muchtar, stadion berkapasitas 38.000 ini akan tetap diprioritaskan sebagai markas tim kebanggaan bobotoh selama bergulirnya Liga 1.

Ia pun membeberkan alasannya menyiapkan stadion yang terletak di kawasan Gedebage ini sebagai kandang Maung Bandung selain Stadion Si Jalak Harupat, Kabupaten Bandung.

"Kalau melihat kondisi sekarang, ya lebih baik di GBLA saja. Sebab, di sini lebih besar (kapasitasnya)," kata manajer yang akrab diasapa Pak Haji ini.

Umuh menambahkan, untuk pertandingan perdana, pihaknya saat ini terus mempersiapkan segala sesuatunya terutama masalah keamanan.

Apalagi, pada pertandingan menghadapi tim berjuluk Singo Edan ini, rencanannya Presiden Joko Widodo juga akan hadir langsung di Stadion GBLA.

"Pak Kapolda pun akan turun tangan mengamankan pertandingan tersebut. Selain itu, kalau memang nanti Pak Presiden datang, pasti dari kepresidenan juga akan ikut membantu mengamankan jalannya pertandingan," ucapnya.

Selain pengamanan di dalam Stadion, Umuh juga akan memperketat keamanan di sekitar Stadion dan akan mengambil tindakan tegas jika ada pelanggaran. Pasalnya, pada beberapa laga terakhir, banyak aduan mengenai pungutan liar kepada bobotoh.

"Insya Allah itu juga akan kami tertibkan masalah perparkiran. Jadi, nanti tidak akan ada lagi parkir hingga Rp 20 ribu sampai Rp 50 ribu, karena itu pemerasan," tegas Umuh.

Gambar 4.30 Teks Asli Artikel-2

4.4 Pembahasan

Evaluasi performa sistem peringkasan teks otomatis dalam penelitian ini menggunakan metrik ROUGE-1, ROUGE-2, dan ROUGE-L. Dalam evaluasi ini, *recall* mengukur seberapa banyak elemen dari ringkasan referensi yang berhasil tercakup dalam ringkasan yang dihasilkan sistem. *Precision* mengukur seberapa banyak elemen dalam ringkasan sistem yang sesuai dengan ringkasan referensi. Sedangkan *f1-score* merupakan nilai rata-rata harmonis dari *recall* dan *precision* yang memberikan keseimbangan antara kedua metrik tersebut. Ketiga nilai metrik ini memiliki rentang antara 0 hingga 1, di mana nilai yang lebih tinggi menunjukkan kinerja sistem yang lebih baik.

Dari hasil uji coba kedua metode, terlihat pola penurunan nilai maksimal seiring dengan meningkatnya tingkat kompresi dari 30% hingga 70%. Pada metode *K-Means Clustering* nilai maksimal ROUGE-1 menurun dari 0,974359 pada kompresi 30% menjadi 0,84375 pada kompresi 70%. Nilai maksimal ROUGE-2 menurun dari 0,973913 pada kompresi 30% menjadi 0,797468 pada kompresi 70%. Begitu juga dengan nilai maksimal ROUGE-L yang menurun dari 0,974359 pada kompresi 30% menjadi 0,84375 pada kompresi 70%.

Sementara pada metode *Textrank*, nilai maksimal ROUGE-1 menurun dari 1 pada kompresi 30% menjadi 0,835821 pada kompresi 70%. Nilai maksimal ROUGE-2 menurun dari 1 pada kompresi 30% menjadi 0,818182 pada kompresi 70%. Begitu juga nilai maksimal ROUGE-L yang menurun dari 1 pada kompresi 30% menjadi 0,835821 pada kompresi 70%. Penurunan ini disebabkan oleh ketidaksetaraan panjang antara ringkasan sistem dan ringkasan manual referensi. Pada tingkat kompresi tinggi (seperti 70%), ringkasan yang dihasilkan sistem menjadi jauh lebih pendek dibandingkan dengan ringkasan manual yang digunakan sebagai standar evaluasi, sehingga secara alami mengurangi kemungkinan *overlap* yang terdeteksi oleh metrik ROUGE.

Berdasarkan hasil uji coba yang telah dilakukan, dapat dilakukan analisis terhadap performa sistem pada kedua metode peringkasan dengan berbagai tingkat kompresi. Pada Tabel 4.11 dan Tabel 4.12 ditampilkan rata-rata hasil evaluasi ROUGE yang membandingkan performa sistem pada 5 skenario tingkat kompresi yaitu 30%, 40%, 50%, 60%, dan 70%, baik untuk metode *K-Means Clustering* maupun *TextRank*. Hasil evaluasi ini mencakup nilai rata-rata *recall*, *precision*, dan

f1-score untuk kelima tingkat kompresi yang diujikan pada masing-masing metode, sehingga dapat dianalisis bagaimana performa dari kedua metode tersebut pada setiap skenario kompresi.

Tabel 4.11 Rata-Rata Hasil Evaluasi ROUGE Menggunakan *K-Means Clustering*

Kompresi	**ROUGE-1**			**ROUGE-2**			**ROUGE-L**		
	Rata-rata *Recall*	**Rata-rata *Precision***	**Rata-rata *F1-Score***	**Rata-rata *Recall***	**Rata-rata *Precision***	**Rata-rata *F1-Score***	**Rata-rata *Recall***	**Rata-rata *Precision***	**Rata-rata *F1-Score***
30%	0.593	0.534	0.550	0.493	0.443	0.456	0.539	0.486	0.500
40%	0.719	0.452	0.546	0.608	0.377	0.458	0.650	0.407	0.493
50%	0.839	0.407	0.542	0.742	0.355	0.474	0.780	0.376	0.500
60%	0.905	0.375	0.524	0.823	0.336	0.472	0.855	0.352	0.494
70%	0.940	0.339	0.494	0.870	0.309	0.452	0.898	0.323	0.471

Berdasarkan hasil rata-rata evaluasi ROUGE pada Tabel 4.11, untuk metode *K-Means Clustering* tingkat kompresi 30% menunjukkan performa terbaik dengan nilai *f1-score* ROUGE-1 sebesar 0.550, diikuti oleh tingkat kompresi 40% dengan nilai *f1-score* 0.546. Untuk ROUGE-2, nilai *f1-score* tertinggi diperoleh pada tingkat kompresi 50% dengan nilai 0.474, sedangkan untuk ROUGE-L, nilai *f1-score* tertinggi terdapat pada tingkat kompresi 30% dan 50% dengan nilai yang sama yaitu 0.500. Secara keseluruhan, seiring dengan peningkatan tingkat kompresi hingga 70%, terlihat pola yang konsisten di mana nilai *recall* terus meningkat dari 0.593 menjadi 0.940 untuk ROUGE-1, dari 0.493 menjadi 0.870 untuk ROUGE-2, dan dari 0.539 menjadi 0.898 untuk ROUGE-L, sementara nilai *precision*

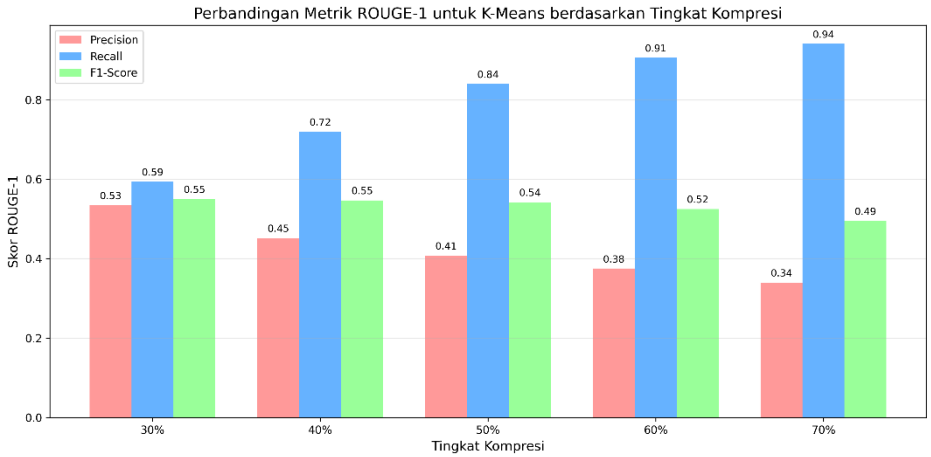
mengalami penurunan dari 0.534 menjadi 0.339 untuk ROUGE-1, dari 0.443 menjadi 0.309 untuk ROUGE-2, dan dari 0.486 menjadi 0.323 untuk ROUGE-L.

Tabel 4.12 Rata-Rata Hasil Evaluasi ROUGE Menggunakan *TextRank*

Kompresi	**ROUGE-1**			**ROUGE-2**			**ROUGE-L**		
	Rata-rata *Recall*	**Rata-rata *Precision***	**Rata-rata *F1-Score***	**Rata-rata *Recall***	**Rata-rata *Precision***	**Rata-rata *F1-Score***	**Rata-rata *Recall***	**Rata-rata *Precision***	**Rata-rata *F1-Score***
30%	0.664	0.547	0.589	0.561	0.459	0.495	0.598	0.492	0.530
40%	0.773	0.473	0.578	0.675	0.408	0.501	0.712	0.433	0.531
50%	0.855	0.414	0.551	0.772	0.369	0.493	0.807	0.389	0.519
60%	0.901	0.375	0.524	0.832	0.341	0.479	0.864	0.359	0.502
70%	0.936	0.341	0.496	0.881	0.316	0.461	0.911	0.331	0.481

Untuk metode *Textrank* terlihat pada Tabel 4.12 performa terbaik untuk ROUGE-1 dicapai pada tingkat kompresi 30% dengan nilai *f1-score* 0.589, diikuti oleh tingkat kompresi 40% dengan nilai *f1-score* 0.578. Untuk ROUGE-2, performa terbaik dicapai pada tingkat kompresi 40% dengan nilai *f1-score* 0.501, sementara untuk ROUGE-L, performa terbaik juga dicapai pada tingkat kompresi 40% dengan nilai *f1-score* 0.531. Pola serupa juga terlihat dimana nilai *recall* terus meningkat dari 0.664 menjadi 0.936 untuk ROUGE-1, dari 0.561 menjadi 0.881 untuk ROUGE-2, dan dari 0.598 menjadi 0.911 untuk ROUGE-L, sedangkan nilai *precision* mengalami penurunan dari 0.547 menjadi 0.341 untuk ROUGE-1, dari 0.459 menjadi 0.316 untuk ROUGE-2, dan dari 0.492 menjadi 0.331 untuk ROUGE-L seiring bertambahnya tingkat kompresi hingga 70%.

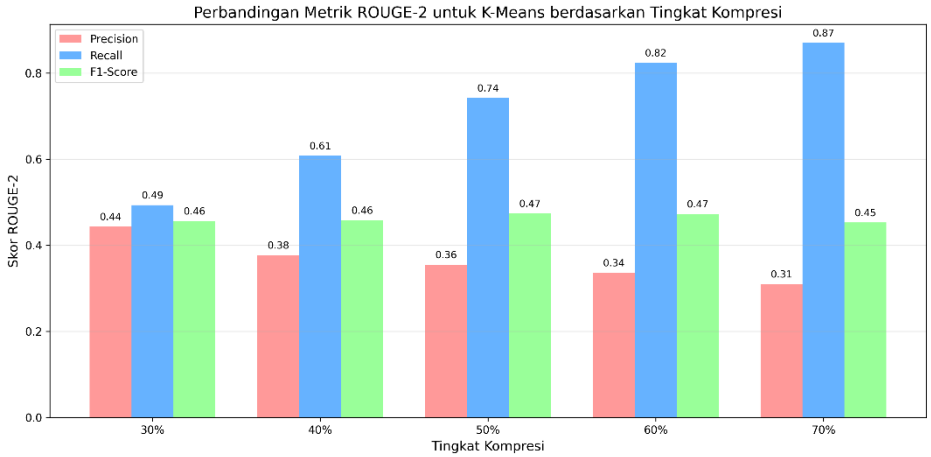
Hasil ini memberikan gambaran bahwa peningkatan tingkat kompresi berbanding lurus dengan peningkatan nilai *recall* namun berbanding terbalik dengan nilai *precision*. Hal ini terjadi karena semakin tinggi tingkat kompresi menyebabkan semakin banyak kata-kata dari ringkasan referensi yang tercakup dalam ringkasan sistem, namun pada saat yang sama juga meningkatkan kemungkinan masuknya kata-kata yang kurang relevan.



Gambar 4.31 Skor Rata-Rata ROUGE-1 *K-Means Clustering*

Visualisasi pada Gambar 4.31 memberikan gambaran tentang pengaruh tingkat kompresi terhadap kualitas ringkasan yang dihasilkan oleh metode *K-Means Clustering* berdasarkan metrik ROUGE-1. Grafik perbandingan metrik ROUGE-1 menunjukkan bahwa seiring meningkatnya tingkat kompresi dari 30% hingga 70%, terjadi peningkatan yang signifikan pada nilai *recall* dari 0.59 pada kompresi 30% hingga mencapai 0.94 pada kompresi 70%. Namun, di sisi lain terlihat penurunan pada nilai *precision* dari 0.53 pada kompresi 30% hingga 0.34 pada kompresi 70%. Nilai *f1-score* tertinggi dicapai pada tingkat kompresi 30% dengan nilai 0.55,

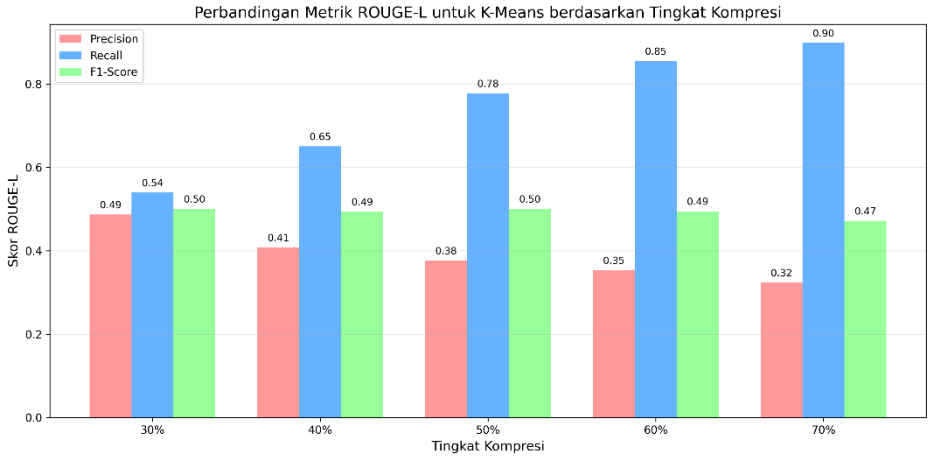
kemudian menurun secara bertahap menjadi 0.54 pada kompresi 50%, 0.52 pada kompresi 60%, hingga 0.49 pada kompresi 70%.



Gambar 4.32 Skor Rata-Rata ROUGE-2 *K-Means Clustering*

Pada Gambar 4.32 menampilkan perbandingan metrik ROUGE-2 untuk metode *K-Means Clustering* berdasarkan tingkat kompresi. ROUGE-2 mengukur kecocokan *bigram* (pasangan kata berurutan) antara ringkasan sistem dan ringkasan referensi, memberikan wawasan tentang koherensi kalimat dalam ringkasan yang dihasilkan. Grafik menunjukkan pola yang serupa dengan ROUGE-1, dimana nilai *recall* mengalami peningkatan signifikan dari 0.49 pada kompresi 30% hingga 0.87 pada kompresi 70%. Sementara itu, nilai *precision* menunjukkan penurunan dari 0.44 pada kompresi 30% hingga 0.31 pada kompresi 70%. Dapat dilihat untuk nilai *f1-score* pada metrik ROUGE-2 ini relatif stabil di kisaran 0.46-0.47 untuk tingkat kompresi 30% hingga 60%, dengan nilai tertinggi 0.47 dicapai pada kompresi 50%, sebelum sedikit menurun menjadi 0.45 pada kompresi 70%. Hal ini mengindikasikan bahwa untuk kecocokan *bigram*, metode *K-Means Clustering*

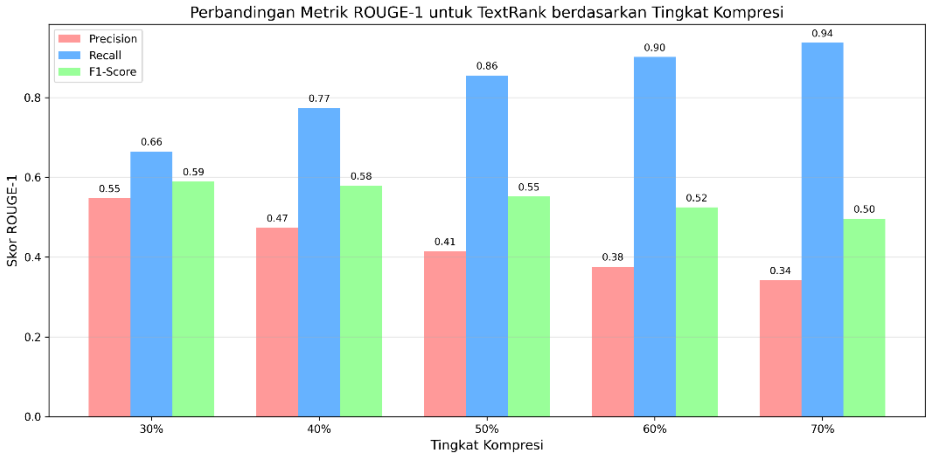
mampu mempertahankan keseimbangan antara *precision* dan *recall* yang relatif stabil pada berbagai tingkat kompresi.



Gambar 4.33 Skor Rata-Rata ROUGE-L *K-Means Clustering*

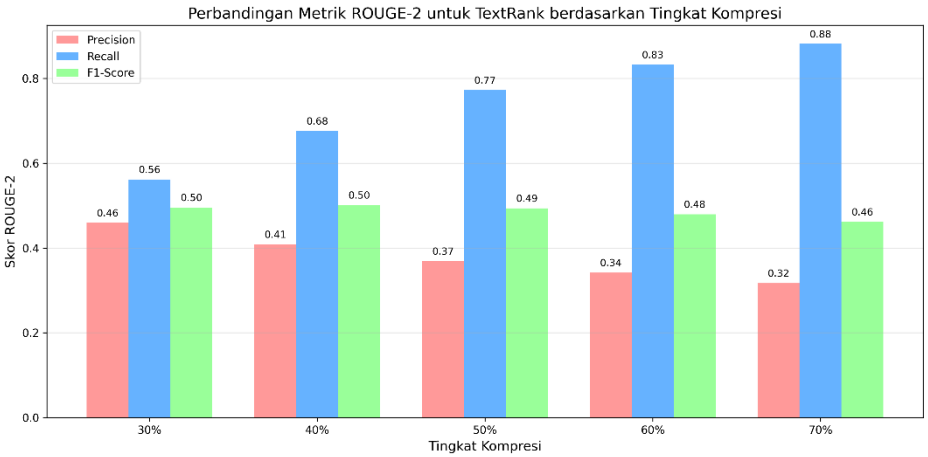
Visualisasi pada Gambar 4.33 menampilkan perbandingan metrik ROUGE-L untuk metode *K-Means Clustering* berdasarkan tingkat kompresi. ROUGE-L mengukur *Longest Common Subsequence* (LCS) antara ringkasan sistem dan ringkasan referensi, memberikan wawasan tentang struktur kalimat dan fleksibilitas urutan kata dalam ringkasan. Grafik menunjukkan bahwa nilai *recall* terus meningkat secara signifikan dari 0.54 pada kompresi 30% hingga mencapai 0.90 pada kompresi 70%. Sementara itu, nilai *precision* mengalami penurunan konsisten dari 0.49 pada kompresi 30% hingga 0.32 pada kompresi 70%. Untuk nilai *f1-score* pada metrik ROUGE-L, terlihat cukup stabil di sekitar 0.49-0.50 pada tingkat kompresi 30% hingga 60%, dengan nilai tertinggi 0.50 dicapai pada kompresi 30% dan 50%, sebelum mengalami sedikit penurunan menjadi 0.47 pada kompresi 70%. Pola ini mengindikasikan bahwa dalam hal struktur kalimat dan urutan kata, metode

K-Means Clustering mampu mempertahankan keseimbangan yang relatif stabil antara *precision* dan *recall* pada berbagai tingkat kompresi.



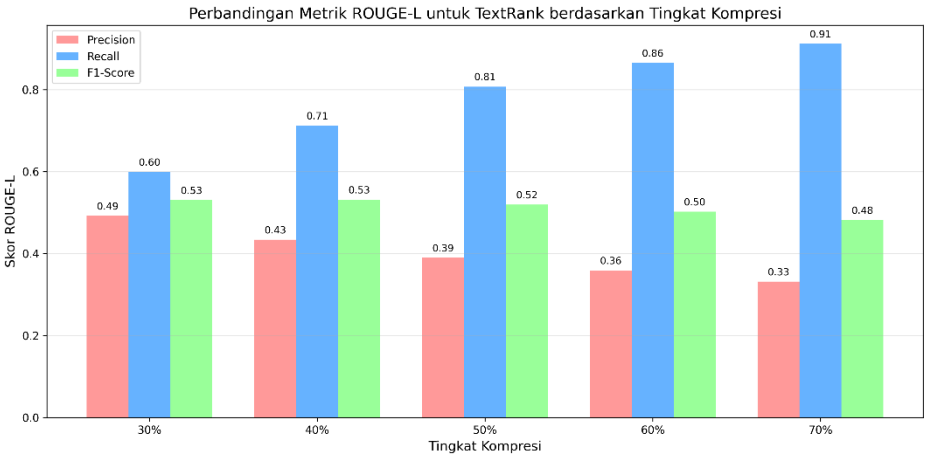
Gambar 4.34 Skor Rata-Rata ROUGE-1 *Textrank*

Visualisasi pada Gambar 4.34 menampilkan perbandingan metrik ROUGE-1 untuk metode *TextRank* berdasarkan tingkat kompresi. Grafik ini memperlihatkan pola yang serupa dengan metode *K-Means Clustering*, dimana terjadi peningkatan nilai *recall* secara signifikan dari 0.66 pada kompresi 30% hingga 0.94 pada kompresi 70%. Di sisi lain, nilai *precision* mengalami penurunan konsisten dari 0.55 pada kompresi 30% hingga 0.34 pada kompresi 70%. Untuk nilai *f1-score*, terlihat penurunan bertahap seiring dengan meningkatnya tingkat kompresi, yakni dari nilai tertinggi 0.59 pada kompresi 30%, menurun menjadi 0.58 pada kompresi 40%, 0.55 pada kompresi 50%, 0.52 pada kompresi 60%, hingga 0.50 pada kompresi 70%. Ini menunjukkan bahwa untuk metrik ROUGE-1, metode *Textrank* memberikan keseimbangan terbaik antara *precision* dan *recall* pada tingkat kompresi yang lebih rendah, khususnya pada kompresi 30%.



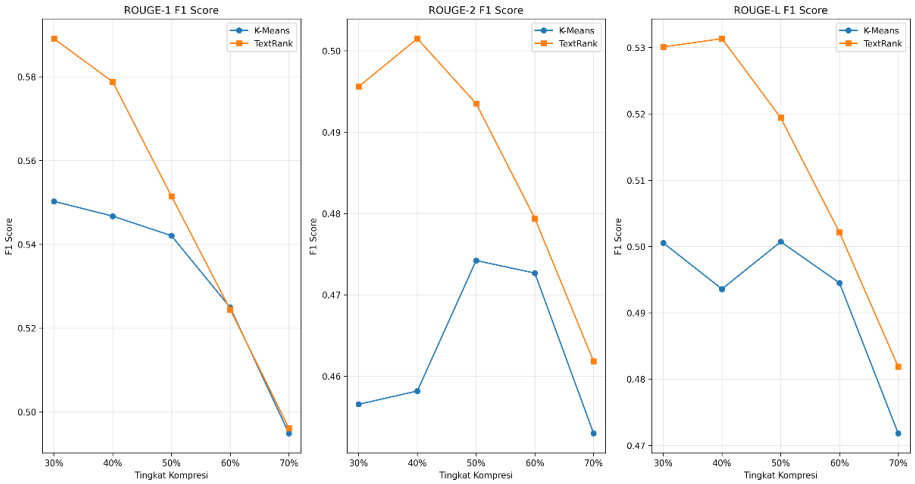
Gambar 4.35 Skor Rata-Rata ROUGE-2 *Textrank*

Visualisasi pada Gambar 4.35 menampilkan perbandingan metrik ROUGE-2 untuk metode *Textrank* berdasarkan tingkat kompresi, dimana nilai *recall* meningkat signifikan dari 0.56 pada kompresi 30% hingga 0.88 pada kompresi 70%, sementara nilai *precision* menurun dari 0.46 menjadi 0.32, dengan nilai *f1-score* yang relatif stabil pada awalnya (0.50 pada kompresi 30% dan 40%) sebelum menurun bertahap menjadi 0.49, 0.48, dan 0.46 pada kompresi yang lebih tinggi, menunjukkan bahwa metode ini memberikan keseimbangan terbaik antara *precision* dan *recall* pada tingkat kompresi yang lebih rendah.



Gambar 4.36 Skor Rata-Rata ROUGE-L *Textrank*

Visualisasi pada Gambar 4.36 menampilkan perbandingan metrik ROUGE-L untuk metode *Textrank* berdasarkan tingkat kompresi, dimana nilai *recall* meningkat secara konsisten dari 0.60 pada kompresi 30% hingga 0.91 pada kompresi 70%, sementara nilai *precision* menurun dari 0.49 menjadi 0.33, dengan nilai *f1-score* yang cenderung menurun dari 0.53 pada kompresi 30% dan 40%, menjadi 0.52 pada kompresi 50%, 0.50 pada kompresi 60%, dan 0.48 pada kompresi 70%, menunjukkan bahwa untuk metrik ini metode *Textrank* memberikan performa terbaik pada tingkat kompresi yang lebih rendah.



Gambar 4.37 Perbandingan Rata-Rata *F1-Score*

Visualisasi pada Gambar 4.37 menunjukkan perbandingan nilai *f1-score* antara metode *K-Means Clustering* dan *Textrank* pada berbagai tingkat kompresi untuk ketiga metrik ROUGE. Dari ketiga grafik tersebut, terlihat jelas bahwa metode *Textrank* (garis oranye) secara konsisten mengungguli metode *K-Means Clustering* (garis biru) pada tingkat kompresi 30%, 40%, dan 50% di semua metrik. Pada ROUGE-1, *Textrank* mencapai *f1-score* tertinggi sekitar 0.59 pada kompresi 30% dan menurun secara konsisten, sementara *K-Means* lebih stabil di rentang 0.54-0.55 sebelum menurun pada kompresi 60% dan 70%. Untuk ROUGE-2, *Textrank* menunjukkan performa tertinggi di kompresi 40% (sekitar 0.50), sedangkan *K-Means* justru mencapai puncaknya di kompresi 50%-60%. Pada ROUGE-L, pola serupa terlihat dengan *Textrank* unggul di kompresi rendah, dengan kedua metode cenderung mirip pada tingkat kompresi tinggi (70%) di sekitar 0.47-0.48.

Hasil yang diperoleh dari kelima skenario dengan tingkat kompresi yang berbeda pada kedua metode menunjukkan peningkatan signifikan dibandingkan dengan penelitian sebelumnya yang menggunakan dataset yang sama. Pada penelitian (Purnama & Utami, 2023) yang menggunakan metode *Text To Text Transfer Transformer* (T5) dengan dataset IndoSum menghasilkan nilai ROUGE-1 terbaik sebesar 0.17568. Sementara penelitian yang dilakukan oleh (Saputra & Maki, 2021) dengan menggunakan metode *Long Short-Term Memory* (LSTM) memperoleh hasil terbaik dengan nilai ROUGE-1 sebesar 0.13846.

Berdasarkan implementasi yang dilakukan, sistem berhasil melakukan peringkasan teks otomatis untuk berita berbahasa Indonesia dengan

membandingkan kinerja dua metode berbeda yaitu *K-Means Clustering* dan *Textrank*. Hasil analisis perbandingan menggunakan metrik ROUGE-1, ROUGE-2, dan ROUGE-L menunjukkan bahwa *Textrank* secara konsisten unggul pada tingkat kompresi rendah hingga menengah (30%-50%) di ketiga metrik dengan nilai *f1-score* yang lebih tinggi. *Textrank* mencapai performa terbaik pada kompresi 30% untuk ROUGE-1 (0,59) dan ROUGE-L (0,53), serta pada kompresi 40% untuk ROUGE-2 (0,50). Sementara itu, *K-Means Clustering* menunjukkan pola yang menarik pada ROUGE-2 dengan kinerja justru meningkat pada kompresi 50%-60%. Pada tingkat kompresi yang lebih tinggi (60%-70%), kedua metode menunjukkan performa yang cenderung mirip, terutama pada kompresi 70% di mana perbedaan keduanya sangat kecil. Penelitian perbandingan ini mengindikasikan bahwa *Textrank* lebih efektif dalam mempertahankan keseimbangan antara cakupan dan ketepatan informasi pada tingkat kompresi rendah, sementara keunggulan tersebut berkurang pada tingkat kompresi yang lebih tinggi untuk dokumen berbahasa Indonesia.

4.5 Integrasi Islam

Integrasi Islam dalam penelitian ini menunjukkan bagaimana pengembangan teknologi dapat sejalan dengan nilai-nilai Islam. Pengembangan sistem peringkasan otomatis tidak hanya berfokus pada aspek teknis semata, tetapi juga mempertimbangkan bagaimana teknologi ini dapat memberikan manfaat sesuai dengan ajaran Islam. Dalam Islam, setiap perbuatan dan pengembangan ilmu pengetahuan hendaknya memiliki nilai ibadah dan memberikan kemanfaatan bagi sesama. Pengembangan teknologi informasi, khususnya dalam bidang peringkasan

teks otomatis, dapat menjadi sarana untuk mengimplementasikan nilai-nilai Islam dalam kehidupan modern.

4.5.1 Muamalah Ma'a Allah (Hubungan Dengan Allah)

Dalam pengembangan sistem peringkasan ini, terdapat nilai-nilai yang mencerminkan ketaatan dan hubungan dengan Allah SWT. Ketaatan tersebut dilandaskan pada kewajiban kita sebagai manusia untuk senantiasa menaati perintah Allah SWT. Hal ini sejalan dengan ayat Al-Qur'an yang menjelaskan mengenai sifat *tabayyun* sesuai dengan firman Allah SWT dalam Al-Qur'an Surah Al-Hujurat ayat 6:

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْٓا اِنْ جَاۤءَكُمْ فَاسِقٌۢ بِنَبَاٍ فَتَبَيَّنُوْٓا اَنْ تُصِيْبُوْا قَوْمًاۢ بِجَهَالَةٍ فَتُصْبِحُوْا عَلٰى مَا فَعَلْتُمْ نٰدِمِيْنَ

"*Wahai orang-orang yang beriman, jika seorang fasik datang kepadamu membawa berita penting, maka telitilah kebenarannya agar kamu tidak mencelakakan suatu kaum karena ketidaktahuan(-mu) yang berakibat kamu menyesali perbuatanmu itu.*" *(QS.Al-Hujurat:6)*

Ayat tersebut dengan jelas menjadi landasan bagi peneliti untuk menerapkan sifat *tabayyun* dalam pengembangan sistem peringkasan teks ini. *Tabayyun* yang berarti meneliti dan mengecek kebenaran informasi, merupakan bentuk ketaatan kepada Allah SWT dalam memastikan bahwa sistem yang dikembangkan tidak menyesatkan pengguna dan memberikan informasi peringkasan yang jelas dan benar. Dengan berlandaskan pada perintah Allah untuk tolong-menolong dalam kebajikan dan takwa, penelitian ini menerapkan prinsip kehati-hatian dan verifikasi pada setiap tahap pengembangan sistemnya secara keseluruhan (Lajnah Pentashihan Mushaf Al-Qur'an, 2019).

Penerapan sifat *tabayyun* dalam sistem ini diwujudkan melalui proses penelitian dan pengujian yang menyeluruh untuk memastikan bahwa hasil peringkasan yang dihasilkan benar dan tidak menyimpang dari informasi aslinya. Setiap komponen sistem dirancang dengan memperhatikan aspek keakuratan dan kebenaran informasi sebagai wujud ketaatan kepada Allah SWT. Dengan menjadikan ketaatan kepada Allah sebagai landasan utama, pengembangan sistem ini tidak hanya bertujuan untuk kemajuan teknologi semata, tetapi juga sebagai sarana untuk melaksanakan perintah Allah dalam menyebarkan informasi yang benar dan bermanfaat.

Adapun sifat *tabayyun* yang diterapkan dalam pengembangan sistem peringkasan teks ini sejalan dengan hadis yang diriwayatkan oleh Abu Daud, dimana Rasulullah SAW bersabda:

التَّأَنِّي مِنَ اللهِ، وَالْعَجَلَةُ مِنَ الشَّيْطَانِ

"Dari Abu Darda, ia berkata: Rasulullah shallallahu 'alaihi wa sallam bersabda: 'Kehati-hatian (tidak tergesa-gesa) itu dari Allah, dan ketergesa-gesaan itu dari syaitan.'' (HR. Abu Daud dan Tirmidzi)

Hadis ini menunjukkan pentingnya sikap hati-hati dan tidak terburu-buru dalam memproses dan menyampaikan informasi. Dengan menerapkan konsep *tabayyun* dalam sistem peringkasan teks, penelitian ini berupaya memastikan bahwa teknologi yang dikembangkan tidak hanya efisien tetapi juga dapat dipercaya keakuratannya. Proses verifikasi yang menyeluruh dilakukan pada setiap komponen sistem, hingga evaluasi hasil peringkasan. Hal ini mencerminkan implementasi dari nilai-nilai Islam yang mengajarkan untuk selalu berhati-hati dan

memverifikasi kebenaran informasi sebelum meneruskannya kepada orang lain (Ustadz Said Yai Ardiansyah Lc MA, 2023).

4.5.2 Muamalah Ma'a An-Naas (Hubungan Dengan Manusia)

Dalam konteks hubungan dengan sesama manusia, pengembangan sistem peringkasan mencerminkan upaya untuk mempermudah dan membantu sesama dalam memahami informasi. Hal ini sejalan dengan Al-Quran Surah Al-Maidah ayat 2:

وَتَعَاوَنُوْا عَلَى الْبِرِّ وَالتَّقْوٰىۖ وَلَا تَعَاوَنُوْا عَلَى الْاِثْمِ وَالْعُدْوَانِۖ

"Dan tolong-menolonglah kamu dalam (mengerjakan) kebajikan dan takwa, dan jangan tolong-menolong dalam berbuat dosa dan permusuhan." (QS.Al-Maidah:2)

Ayat tersebut menjelaskan tentang bagaimana seharusnya kita berinteraksi dengan sesama manusia harus saling tolong menolong dalam berbuat kebajikan. Islam mengajarkan bahwa nilai seseorang tidak hanya diukur dari ibadah ritualnya, tetapi juga dari seberapa besar manfaat yang dapat dia berikan kepada orang lain. Sistem peringkasan yang dikembangkan merupakan implementasi dari ayat ini, di mana teknologi digunakan sebagai sarana untuk memberikan manfaat kepada masyarakat dalam memahami dan mengolah informasi dengan lebih efisien (Lajnah Pentashihan Mushaf Al-Qur'an, 2019).

Nilai-nilai muamalah dalam pengembangan sistem ini juga tercermin dari bagaimana teknologi dapat membantu mengefisienkan waktu dan tenaga pengguna. Dalam konteks memberikan manfaat kepada sesama, sistem ini diharapkan dapat membantu pengguna untuk mendapatkan intisari informasi dengan lebih cepat,

sehingga mereka dapat mengalokasikan waktu dan energi mereka untuk hal-hal produktif lainnya. Ini merupakan bentuk nyata dari upaya untuk menjadi pribadi yang bermanfaat bagi sesama sebagaimana yang dianjurkan dalam ayat tersebut.

Adapun sifat tolong menolong dan bermanfaat untuk sesama manusia yang diterapkan dalam pengembangan sistem peringkasan teks ini sejalan dengan hadis yang diriwayatkan oleh Ahmad, dimana Rasulullah SAW bersabda:

عَنْ جَابِرٍ رَضِيَ اللهُ عَنْهُ، قَالَ: قَالَ رَسُولُ اللَّهِ صَلَّى اللهُ عَلَيْهِ وَسَلَّمَ: "الْمُؤْمِنُ يَأْلَفُ وَيُؤْلَفُ، وَلَا خَيْرَ فِيمَنْ لَا يَأْلَفُ وَلَا يُؤْلَفُ، وَخَيْرُ النَّاسِ أَنْفَعُهُمْ لِلنَّاسِ

"Dari Jabir radhiyallahu 'anhu, ia berkata: Rasulullah shallallahu 'alaihi wa sallam bersabda: 'Seorang mukmin itu bersikap ramah dan tidak ada kebaikan bagi seseorang yang tidak bersikap ramah. Dan sebaik-baik manusia adalah yang paling bermanfaat bagi manusia lain.'" (HR. Ahmad dan Thabrani)

Hadis ini menegaskan bahwa memberikan manfaat kepada orang lain merupakan salah satu ciri dari manusia terbaik. Dalam konteks pengembangan sistem peringkasan teks, upaya untuk menciptakan teknologi yang memudahkan orang dalam memahami informasi merupakan bentuk nyata dari prinsip memberi manfaat kepada sesama. Dengan mengembangkan sistem yang mampu meringkas teks secara efektif dan akurat, penelitian ini berusaha memberikan kontribusi nyata dalam membantu mengurangi kesulitan yang dihadapi banyak orang ketika berhadapan dengan dokumen yang panjang dan kompleks (Detik Hikmah, 2024).

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil uji coba pada 2500 dokumen dataset IndoSum, peringkasan teks berbahasa Indonesia menunjukkan performa yang bervariasi di ketiga metrik ROUGE. Performa optimal terlihat pada tingkat kompresi 30%-50% dikarenakan pada rentang ini menyediakan keseimbangan terbaik antara cakupan informasi dan ketepatan. Metode *Textrank* mengungguli *K-Means Clustering* pada tingkat kompresi 30% dengan nilai *f1-score* ROUGE-1 sekitar 0.59 dibandingkan 0.55, *f1-score* ROUGE-2 sekitar 0.50 dibandingkan 0.46, dan *f1-score* ROUGE-L sekitar 0.53 dibandingkan 0.50. Kedua metode menunjukkan kecenderungan yang sama dimana seiring meningkatnya tingkat kompresi dari 30% hingga 70%, nilai *recall* meningkat sedangkan *precision* menurun.

Sesuai dengan tujuan penelitian untuk membandingkan kinerja kedua metode, hasil evaluasi menunjukkan bahwa *Textrank* secara konsisten memberikan nilai *f1-score* yang lebih tinggi pada tingkat kompresi 30% hingga 50% untuk ketiga metrik ROUGE. Pada ROUGE-2, *K-Means Clustering* menunjukkan pola peningkatan performa pada kompresi 50%-60%. Sementara itu, pada tingkat kompresi 60% hingga 70%, kedua metode menunjukkan performa yang cenderung sama, dengan perbedaan kecil terutama pada kompresi 70%. Berdasarkan temuan ini, dapat disimpulkan bahwa metode *Textrank* lebih sesuai digunakan untuk peringkasan dengan tingkat kompresi rendah hingga menengah (30%-50%),

sementara metode *K-Means Clustering* dapat menjadi alternatif yang layak dipertimbangkan pada tingkat kompresi yang lebih tinggi, terutama untuk aplikasi yang mengutamakan struktur kalimat yang diukur dengan ROUGE-2.

5.2 Saran

Peneliti menyadari bahwa penelitian ini memiliki beberapa kekurangan untuk mencapai hasil yang optimal dalam pembuatan sistem peringkasan teks otomatis dalam Bahasa Indonesia. Diharapkan saran ini dapat menjadi rekomendasi untuk peneliti-peneliti selanjutnya:

1. Penerapan pembelajaran mendalam (*deep learning*) untuk meningkatkan pemahaman semantik dalam proses peringkasan teks berbahasa Indonesia, yang dapat mengatasi keterbatasan pendekatan berbasis vektor TF-IDF dalam menangkap konteks dan makna yang lebih kompleks.
2. Pengembangan teknik *preprocessing* khusus untuk Bahasa Indonesia yang mempertimbangkan struktur pembentukan kata dan pola susunan kalimat yang lebih kompleks, sehingga dapat meningkatkan kualitas representasi kalimat dan pada akhirnya meningkatkan performa kedua metode peringkasan.

DAFTAR PUSTAKA

Abdillah, M. (2024). *Analisis Perbandingan TextRank dan Long Short Term Memory Dalam Peringkasan Teks Berita Bahasa Inggris* [Undergraduate, Universitas Islam Negeri Maulana Malik Ibrahim Malang]. http://etheses.uin-malang.ac.id/65984/

Abdurrohman, A. (2018). *Evaluasi Algoritma Textrank pada Peringkasan Teks Berbahasa Indonesia* [Undergraduate, Universitas Sumatera Utara]. https://repositori.usu.ac.id/handle/123456789/50791

Abdussalam Amrullah, Intam Purnamasari, Betha Nurina Sari, Garno, & Apriade Voutama. (2022). Analisis Cluster Faktor Penunjang Pendidikan Menggunakan Algoritma K-Means (Studi Kasus: Kabupaten Karawang). *Jurnal Informatika Dan Rekayasa Elektronik*, *5*(2), 244–252. https://doi.org/10.36595/jire.v5i2.701

Ahsan, T. (2023). *Peringkasan Teks Multi Dokumen Berita Berbahasa Indonesia Menggunakan FastText Dan K-Means Clustering* [Undergraduate, Universitas Islam Negeri Maulana Malik Ibrahim Malang]. http://etheses.uin-malang.ac.id/52460/

Alamanda, R., Suhery, C., & Brianorman, Y. (2016). Aplikasi Pendeteksi Plagiat Terhadap Karya Tulis Berbasis Web Menggunakan Natural Language Processing dan Algoritma Knuth-Morris-Pratt. *Coding Jurnal Komputer dan Aplikasi*, *4*(1), 33–44. https://doi.org/10.26418/coding.v4i1.13332

Alhadi, M. bin A. (2022). *Jawami' Al-Kalim Nabi*. https://ibihtafsir.id/2022/04/09/jawami-al-kalim-nabi/

Al-Hajiri, R. bin M. bin F. (2018). *Jawami' Al-Kalim "Keindahan Retorika Hadis Nabi Muhammad SAW."* Al-Andalus Group. https://islammessage.org/id/book/3467/Jaw%C4%81mi%E2%80%99-al-Kalim-%C2%AB-Keindahan-Retorika-Hadis-Nabi-Muhammad-SAW%C2%BB

Andriani, D., Indriati, & Furqon, M. T. (2019). Peringkasan Teks Otomatis Pada Artikel Berita Hiburan Berbahasa Indonesia Menggunakan Metode BM25. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(3), Article 3.

Detik Hikmah. (2024). *Sebaik-baik Manusia Adalah yang Bermanfaat bagi Orang Lain*. https://www.detik.com/hikmah/doa-dan-hadits/d-7123193/sebaik-baik-manusia-adalah-yang-bermanfaat-bagi-orang-lain-ini-haditsnya

Fadhila, L. N., & Nuryana, I. K. D. (2024). *Teks Ringkas Otomatis pada Portal Berita CNN Indonesia Menggunakan Algoritma Textrank*. *05*(01).

Halimah, Agustian, S., & Ramadhani, S. (2022). Peringkasan teks otomatis (automated text summarization) pada artikel berbahasa indonesia menggunakan algoritma lexrank. *Jurnal CoSciTech (Computer Science and Information Technology)*, *3*(3), Article 3. https://doi.org/10.37859/coscitech.v3i3.4300

Hernawan, Y. F., Adikara, P. P., & Wihandika, R. C. (2022). Peringkasan Artikel Berbahasa Indonesia Menggunakan TextRank dengan Pembobotan BM25. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *9*(1), 61–68. https://doi.org/10.25126/jtiik.2022913765

Husniah, F., Agustian, S., & Afrianty, I. (2022). Peringkasan Teks Otomatis Artikel Berbahasa Indonesia Menggunakan Algoritma Textrank. *Seminar Nasional Teknoka*, *7*(7), 10.

Khan, R., Qian, Y., & Naeem, S. (2019). Extractive based Text Summarization Using KMeans and TF-IDF. *International Journal of Information Engineering and Electronic Business*, *11*(3), 33–44. https://doi.org/10.5815/ijieeb.2019.03.05

Khatri, C., Singh, G., & Parikh, N. (2018). *Abstractive and Extractive Text Summarization using Document Context Vector and Recurrent Neural Networks* (arXiv:1807.08000; Issue arXiv:1807.08000). arXiv. https://doi.org/10.48550/arXiv.1807.08000

Kurniawan, K., & Louvan, S. (2018). Indosum: A New Benchmark Dataset for Indonesian Text Summarization. *2018 International Conference on Asian Language Processing (IALP)*, 215–220. https://doi.org/10.1109/IALP.2018.8629109

Lajnah Pentashihan Mushaf Al-Qur'an. (2019). *Al-Qur'an Kementerian Agama—Pustaka Lajnah*. https://pustakalajnah.kemenag.go.id/detail/135

Lubis, T. H., & Koto, I. (2020). Diskursus Kebenaran Berita Berdasarkan Undang-Undang Nomor 40 Tahun 1999 Tentang Pers Dan Kode Etik Jurnalistik. *DE LEGA LATA: Jurnal Ilmu Hukum*, *5*(2), Article 2. https://doi.org/10.30596/dll.v5i2.4169

Mutlu, B., Sezer, E. A., & Akcayol, M. A. (2020). Candidate sentence selection for extractive text summarization. *Information Processing & Management*, *57*(6), 102359. https://doi.org/10.1016/j.ipm.2020.102359

Nagalavi, D., & Hanumanthappa, M. (2019). The NLP Techniques for Automatic Multi-article News Summarization Based on Abstract Meaning Representation. In V. S. Rathore, M. Worring, D. K. Mishra, A. Joshi, & S. Maheshwari (Eds.), *Emerging Trends in Expert Applications and Security* (Vol. 841, pp. 253–260). Springer Singapore. https://doi.org/10.1007/978-981-13-2285-3_31

Prabowo, D. A., Fauzi, M. A., & Sari, Y. A. (2017). Peringkasan Teks Ekstraktif Kepustakaan Ilmu Komputer Bahasa Indonesia Menggunakan Metode Normalized Google Distance dan K-means. *JPTIIK*, *1*(12), 1697–1707.

Purnama, I. N., & Utami, N. N. W. (2023). Implementasi Peringkas Dokumen Berbahasa Indonesia Menggunakan Metode Text To Text Transfer Transformer (T5). *Program Studi Sistem Informasi*, 381–391.

Robiyanto, R., Nugraha, N., & Apriatna, I. (2019). Peringkasan Teks Otomatis Berita Menggunakan Metode Maximum Marginal Relevance. *JEJARING : Jurnal Teknologi dan Manajemen Informatika*, *4*(1), Article 1. https://doi.org/10.25134/jejaring.v4i1.6712

Ronaning Roem, E., & Vanisya, W. (2024). Transformation in the digital era: Optimising social media for news coverage. *Jurnal Studi Komunikasi (Indonesian Journal of Communications Studies)*, *8*(3), 675–684. https://doi.org/10.25139/jsk.v8i3.8477

Samosir, F. V. P., Toba, H., & Ayub, M. (2022). BESKlus: BERT Extractive Summarization with K-Means Clustering in Scientific Paper. *Jurnal Teknik Informatika dan Sistem Informasi*, *8*(1). https://doi.org/10.28932/jutisi.v8i1.4474

Saputra, M. A., & Maki, W. F. A. (2021). Peringkas Teks Otomatis Bahasa Indonesia secara Abstraktif menggunakan Metode Long Short-Term Memory. *eProceedings of Engineering*, *8*(2), Article 2.

Shetty, K., & Kallimani, J. S. (2017). Automatic extractive text summarization using K-means clustering. *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, 1–9. https://doi.org/10.1109/ICEECCOT.2017.8284627

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Suputra, I. (2017). Peringkasan teks otomatis untuk dokumen bahasa Bali berbasis metode ektraktif. *Jurnal Ilmiah Ilmu Komputer*, *10*(1), Article 1.

UIN Sayyid Ali Rahmatullah Tulungagung. (2023). *Telaah Tafsir Al Muyassar Jilid II*. http://repo.uinsatu.ac.id/33488/1/TELAAH%20TAFSIR%20AL-MUYASSAR%20Jilid%20II_2.pdf

Ustadz Said Yai Ardiansyah Lc MA. (2023). *Tergesa-gesa Penyakit Manusia*. https://almanhaj.or.id/45725-tergesa-gesa-penyakit-manusia-2.html#_ftn5

Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *J*, *2*(2), 226–235. https://doi.org/10.3390/j2020016