SKRIPSI

Implementasi *TextRank* pada Peringkasan Teks Otomatis Dokumen Berbahasa Indonesia

Disusun dan diajukan oleh:

A. FIQRAM ABDILLAH ASTRI
D42116507

PROGRAM STUDI SARJANA TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS HASANUDDIN
GOWA
2023

LEMBAR PENGESAHAN SKRIPSI

IMPLEMENTASI TEXTRANK PADA PERINGKASAN TEKS OTOMATIS DOKUMEN BERBAHASA INDONESIA

Disusun dan diajukan oleh

A.FIQRAM ABDILLAH ASTRI

D42116507

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Sarjana Program Studi Teknik Informatika Fakultas Teknik Universitas Hasanuddin pada tanggal 27 Juli 2023 dan dinyatakan telah memenuhi syarat kelulusan.

Menyetujui,

Pembimbing Utama,	Pembimbing Pendamping,
Elly Warni, ST.,MT.	Mukarramah Yusuf, B.Sc., M.Sc., Ph.D
Nip. 198202162008122001	Nip. 198310082012122003

Ketua Program Studi,

Prof. Dr. Ir. Indrabayu, ST., MT., M.Bus.Sys., IPM, ASEAN. Eng.
NIP. 197507162002121004

PERNYATAAN KEASLIAN

Yang bertanda tangan dibawah ini ;

Nama	: A. Fiqram Abdillah Astri
NIM	: D42116507
Program Studi	: Teknik Informatika
Jenjang	: S1

Menyatakan dengan ini bahwa karya tulisan saya berjudul

Implementasi *TextRank* pada Peringkasan Teks Otomatis Dokumen Berbahasa Indonesia

Adalah karya tulisan saya sendiri dan bukan merupakan pengambilan alihan tulisan orang lain dan bahwa skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri.

Semua informasi yang ditulis dalam skripsi yang berasal dari penulis lain telah diberi penghargaan, yakni dengan mengutip sumber dan tahun penerbitannya. Oleh karena itu semua tulisan dalam skripsi ini sepenuhnya menjadi tanggung jawab penulis. Apabila ada pihak manapun yang merasa ada kesamaan judul dan atau hasil temuan dalam skripsi ini, maka penulis siap untuk diklarifikasi dan mempertanggungjawabkan segala resiko.

Segala data dan informasi yang diperoleh selama proses pembuatan skripsi, yang akan dipublikasi oleh Penulis di masa depan harus mendapat persetujuan dari Dosen Pembimbing.

Apabila dikemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan isi skripsi ini hasil karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Gowa, 17 – 07 - 2023

Yang Menyatakan

A.Fiqram Abdillah Astri

ABSTRAK

A. FIQRAM ABDILLAH ASTRI. *Implementasi TextRank Pada Peringkas Teks Otomatis Dokumen Berbahasa Indonesia* (dibimbing oleh Elly Warni, S.T.,M.T dan Mukarramah Yusuf, B.Sc., M.Sc., Ph.D)

Peringkasan teks otomatis adalah proses pengurangan teks yang dilakukan secara otomatis oleh komputer atau algoritma komputasi. Peringkasan teks menunjukkan bahwa proses ini membantu mengurangi teks yang panjang menjadi ringkasan yang lebih singkat tetapi masih mempertahankan informasi penting. Masalah dari penelitian ini adalah bagaimana membuat peringkasan teks ototmatis dokumen artikel website berbahasa Indonesia dengan mengimplementasikan metode *TextRank*. Melalui analisis data teks bahasa Indonesia. Lokasi penelitian berlangsung di Laboratorium Animasi dan Multimedia, Departemen Teknik Informatika, Fakultas Teknik, Universitas Hasanuddin. *Use Case Diagram* Sistem Aplikasi yang dilakukan actor melakukan input teks hingga menghasilkan ringkasan teks. Hasil penelitian menunjukkan bahwa sistem peringkasan teks otomatis menggunakan metode *TextRank* menghasilkan tingkat keakuratan yang tinggi. Skor *ROUGE-1* dengan *Precision* sebesar 1.0, *Recall* sebesar 0.29 dan *F1-Score* sebesar 0.45. Skor *ROUGE-2* dengan *Precision* sebesar 1.0, *Recall* sebesar 0.23 dan *F1-Score* sebesar 0.38. Skor *ROUGE-L* dengan *Precision* sebesar 1.0, *Recall* sebesar 0.29 dan *F1-Score* sebesar 0.45. Hasil ini menunjukkan tingkat kesesuaian yang baik antara ringkasan yang dihasilkan dengan teks asli.

Kata kunci: Peringkasan Teks Otomatis, *TextRank*, *ROUGE*.

ABSTRACT

A. FIQRAM ABDILLAH ASTRI. *Implementasi TextRank Pada Peringkas Teks Otomatis Dokumen Berbahasa Indonesia* (supervised by Elly Warni, S.T.,M.T and Mukarramah Yusuf, B.Sc., M.Sc., Ph.D)

Automatic text summarization is a process of reducing text automatically by a computer or computational algorithm. Text summarization suggests that this process helps reduce long texts into shorter summaries while still retaining important information. The problem of this research is how to make automatic text summarization of Indonesian language website article documents by implementing the TextRank method. Through data analysis of Indonesian text. The location of the research took place at the Animation and Multimedia Laboratory, Department of Informatics Engineering, Faculty of Engineering, Hasanuddin University. Use Case System Diagram The application that the actor performs inputs text to produce a text summary. The results of this research show that the automatic text summarization system using the TextRank method produces a high level of accuracy. ROUGE-1 score with Precision of 1.0, Recall of 0.29 and F1-Score of 0.45. ROUGE-2 score with Precision of 1.0, Recall of 0.23 and F1-Score of 0.38. ROUGE-L score with Precision of 1.0, Recall of 0.29 and F1-Score of 0.45. These results indicate a good degree of correspondence between the resulting summary and the original text.

Keywords: Automated Text Summarization, TextRank, ROUGE.

KATA PENGANTAR

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

Segala puji bagi Allah SWT yang telah memberikan rahmat dan karuniaNya kepada penulis, sehingga penulis dapat menyelesaikan skripsi berjudul **"Implementasi *TextRank* pada Peringkasan Teks Otomatis Dokumen Berbahasa Indonesia"** dengan baik. Shalawat dan salam senantiasa tercurah kepada Rasulullah SAW yang mengantarkan manusia dari zaman kegelapan ke zaman yang terang benderang ini. Penyusunan skripsi ini sebagai salah satu syarat untuk menyelesaikan Program Sarjana (S1) pada Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin.

Penulis menyadari bahwa penulisan ini tidak dapat terselesaikan tanpa dukungan dari berbagai pihak baik morel maupun materil. Oleh karena itu, penulis ingin menyampaikan ucapan terima kasih kepada semua pihak yang telah membantu dalam penyusunan skripsi ini terutama kepada:

1. Kedua orang tua, Drs. H. A. Syamsuddin Tali dan Dra. Hj. Husnaeni Alief, kakak dan adik – adik saya yang telah memberikan dukungan dan motivasi serta doa yang tiada henti-hentinya kepada penulis.
2. Ibu Elly Warni, S.T., M.T., selaku dosen pembimbing I dan Ibu Mukarramah Yusuf, B.Sc., M.Sc., Ph.D., selaku dosen pembimbing II, yang selalu menyediakan waktu, tenaga dan pikirannya yang luar biasa untuk mengarahkan penulis dalam penyusunan Tugas Akhir ini.
3. Bapak Bapak Prof. Dr. Ir. Indrabayu, S.T., M.T., M.Bus.Sys., IPM., ASEAN., Eng., selaku Ketua Departemen Teknik Informatika Fakultas Teknik Univeristas Hasanuddin atas bimbingannya selama masa perkuliahan.
4. Segenap Dosen dan Staff Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin, yang telah banyak membantu penulis selama masa perkuliahan.
5. Kepada Saudara – Saudari IGNITER 16 yang telah berjuang Bersama dari awal sampai akhir penyelesaian skripsi ini, yang selalu memberikan motivasi dan semangat yang luar biasa. Sehingga masa – masa perkuliahan terasa sangat meneyenangkan.
6. Kepada seseorang Wanita yaitu Islamiah yang selalu memberikan semangat kepada penulis agar selalu mengerjakan skripsinya.
7. Semua pihak atas dukungan dan bantuannya yang tidak dapat penulis tuliskan satu persatu.

Penulis menyadari bahwa skripsi ini masih jauh dari sempurna dikarenakan terbatasnya pengalaman dan pengetahuan yang dimiliki penulis. Oleh karena itu, penulis mengharapkan segala bentuk saran serta masukan bahkan kritik yang membangun dari berbagai pihak. Semoga skripsi ini dapat bermanfaat bagi para pembaca dan semua pihak.

Makassar, Januari 2023

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi di era revolusi 4.0 dan *big* data yang eranya adalah penerapan teknologi *modern*. Saat ini terjadi ledakan jumlah data teks dari berbagai sumber. Volume teks ini adalah sumber informasi dan pengetahuan yang tak ternilai yang perlu diringkas secara efektif. Ketersediaan dokumen yang meningkat ini secara tidak langsung membutuhkan peringkasan teks otomatis untuk mengurangi waktu dan tenaga untuk menemukan informasi yang ringkas dan relevan berdasarkan *query*. Peringkasan teks otomatis adalah tugas menghasilkan ringkasan yang ringkas dan lancar tanpa bantuan manusia dengan tetap mempertahankan arti dari dokumen teks asli. Untuk dapat mengetahui informasi penting dari suatu dokumen, pembaca harus meluangkan banyak waktu. Oleh sebab itu, jika dokumen tersebut dapat diringkas oleh suatu sistem tanpa menghilangkan informasi yang penting maka pembaca dapat menghemat waktu. Karena pembaca dapat memahami dan mengetahui informasi penting dari dokumen tersebut tanpa harus membaca isi dokumen secara keseluruhan.

Meringkas adalah mereduksi isi teks dengan tetap menjaga maksud dari teks sebelumnya. Ringkasan yang baik memberikan ringkasan singkat tentang poin-poin penting dalam bagian tertentu. Peringkasan teks telah banyak digunakan dalam berbagai bidang seperti sains, pendidikan, hukum, kedokteran, teknik dan sebagainya. Para peneliti berfokus pada penulisan ringkasan resep dokter, yang terbukti sangat berguna bagi pasien. Demikian pula, artikel ilmiah berbentuk panjang diringkas sehingga pembaca bisa mendapatkan banyak informasi tentang beberapa informasi dalam waktu singkat.

Peringkasan teks otomatis adalah proses pengurangan teks yang dilakukan secara otomatis oleh komputer atau algoritma komputasi. Peringkasan sebuah teks yang dilakukan secara manual oleh manusia tentunya harus membacanya secara keseluruhan untuk mengembangkan pemahaman kita, dan kemudian menulis ringkasan yang berdasarkan poin-poin utamanya. Karena komputer tidak memiliki pengetahuan manusia dan kemampuan bahasa, itu membuat peringkasan teks secara manual menjadi tugas yang terbilang sulit, melelahkan dan membutuhkan waktu yang cukup lama. Meringkas dokumen secara manual oleh manusia, membutuhkan banyak biaya dan waktu apabila dokumen tersebut banyak dan panjang sehingga diperlukan sistem peringkas otomatis (*automatic summarization*) untuk mengatasi banyaknya biaya dan waktu tersebut.

Peringkasan teks otomatis adalah tugas menghasilkan ringkasan ringkas dan lancar tanpa bantuan manusia. Hal ini tentu menjadi masalah dimana ringkasan dibuat dengan tujuan untuk meminimalkan waktu membaca dan memudahkan para pembaca. Terutama dalam teks yang panjang, tidak mudah mendapatkan informasi dari artikel yang panjang dalam waktu yang singkat. Oleh karena itu diperlukan suatu sistem yang dapat meringkas teks secara otomatis untuk

membantu dan mempermudah pembaca artikel ilmiah dalam menghemat waktu membaca dan juga dapat memberikan informasi yang cepat dengan membaca hasil ringkasan (*summary*) artikel dengan memanfaatkan peringkas teks otomatis. Penelitian ini akan membangun peringkas teks otomatis dengan menginplementasikan *TextRank* pada peringkas teks otomatis dokumen bahasa Indonesia. Penelitian pada algoritma peringkas teks otomatis terus berkembang dan dapat diklasifikasikan menjadi beberapa metode. *Riview* jurnal beberapa metode yang di lakukan oleh W Widodo dengan judul *A Comparative Riview Of Extractive Text Summarization In Indonesia Languange* pada tahun (2021) yang membandingkan 6 metode yaitu, *Vector Space Model, Maximum Marginal Relevance, TextTeaser, Graph Convulution Networks, Sentence Scoring* dan *Decision Tree*, dan terakhir *Sentence fusion*. dari beberapa peringkas teks masih jarang untuk artikel jurnal dalam bahasa Indonesia. Hasil dari riview pada 6 metode di atas di dapatkan metode terbaik pada jurnal Sabuna P M, dkk (2017) *Summarizing Indonesian text automatically by using sentence scoring and Decision Tree*. Jadi hasil penelitian yang mengunakan *sentence scoring* TF-IDF menunjukkan pengukuran nilai f-measure tertinggi adalah 0,80 dan rata-rata 0,58. Berdasarkan hasil tersebut dapat disimpulkan hasil peringkasan dokumen menngunakan *sentence scoring* menunjukkan hasil akurasi yang cukup baik untuk document teks. Adapun saran untuk mengembangkan formula perhitungan nilai fitur teks yang lebih baik. Harapanya nilai setiap fitur yang dihasilkan nantinya lebih beragam, sehingga dapat menghasilkan model aturan yng lebih baik. I G M Darmawiguna, dkk (2020) Indonesian sentiment summarization for lecturer learning evaluation by using *TextRank* algorithm. Berdasarkan hasil Algoritma *TextRank* dapat digunakan untuk meringkas opini dengan baik dengan akurasi 82%. Rafael Ferreira, dkk *Assessing sentence scoring techniques for extractive text summarization*, saya mengambil rekomendasi dari jurnal ini yang dimana sentence scoring dapat menggunakan *TextRank*.

Berdasarkan hasil penelitian sebelumnya yang di uraikan di atas, maka pada penelitian ini akan dibangun suatu sistem "IMPLEMENTASI *TEXTRANK* PADA PERINGKAS TEKS OTOMATIS DOKUMEN BERBAHASA INDONESIA" dengan menggunakan algoritma *TextRank* untuk peningkatan akurasi dan hasil yang lebih baik terhadap peringkas teks otomatis, sehingga nantinya penelitian ini dapat mempermudah dan mempersingkat waktu peringkasan pada dokumen teks berbahasa Indonesia.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas dapat dirumuskan masalah penelitian ini adalah bagaimana membuat peringkas teks otomatis dokumen artikel website berbahasa Indonesia dengan mengimplementasikan metode *TextRank*?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah di atas penelitian ini bertujuan untuk membuat peringkasan teks otomatis artikel website berbahasa Indonesia dengan metode *TextRank*.

1.4 Batasan Masalah

Adapun yang menjadi batasan masalah dalam penelitian ini adalah:

1. Variabel domain yang digunakan dalam penelitian ini menggunakan artikel *website* teknologi informasi.
2. Penelitian ini difokuskan pada peringkasan artikel *website* berbahasa Indonesia.

1.5 Manfaat Penelitian

Hasil dari penelitian dapat membantu dalam melakukan peringkasan teks artikel *website* bahasa Indonesia yang dimana tetap terkandung pikiran pokok dari teks.

1.6 Sistematika Penulisan

Untuk memberikan gambaran singkat mengenai isi tulisan secara keseluruhan, maka akan diuraikan beberapa tahapan dari penulisan secara sistematis, yaitu:

BAB I PENDAHULUAN

Bab ini menguraikan secara umum mengenai hal yang menyangkut latar belakang, perumusan masalah dan batasan masalah, tujuan, manfaat dan sistematika penulis.

BAB II TINJAUAN PUSTAKA

Bab ini berisi teori-teori terkait hal-hal yang mendasari dan yang berhubungan dengan visi komputer dan metode-metode yang digunakan pada penelitian ini.

BAB III METODOLOGI PENELITIAN

Bab ini memaparkan tahapan penelitian, waktu dan lokasi penelitian, instrumen penelitian, teknik pengambilan data dan rancangan sistem serta penerapan algoritna dan metode-metode dalam pengolahan data, mulai dari preprocessing hingga menghasilkan prediksi.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi tentang hasil penelitian dan pembahasan terkait pengolahan data yang telah dilakukan yang disertai dengan hasil penelitian.

BAB V PENUTUP

Bab ini berisi kesimpulan yang diperolah dari hasil penelitian yang telah dilakukan serta saran-saran untuk pengembangan sistem lebih lanjut.

BAB II
TINJAUAN PUSTAKA

2.1 *Text summarization*

Text summarization (ringkasan teks) adalah teknik yang digunakan untuk mengurangi teks artikel panjang menjadi potongan yang lebih kecil dan mudah dipahami (Zamzam, 2020). Hal ini memanfaatkan teknik dan prinsip pemrosesan bahasa alami untuk menganalisis artikel dan menghasilkan ringkasan yang lebih pendek namun efektif. Meringkas otomatis melibatkan penggunaan perangkat lunak komputer untuk memadatkan teks dan mempertahankan informasi terpenting dari sumber aslinya. Meringkas data otomatis merupakan komponen penting dalam machine learning dan data mining. Teknologi ini banyak digunakan dalam berbagai industri saat ini, seperti mesin pencari seperti Google yang dapat memberikan ringkasan hasil pencarian. Selain itu, meringkas juga dapat diterapkan pada makalah, koleksi gambar, atau ringkasan film.

Dalam meringkas, tujuan utamanya adalah mencari segmen yang mewakili setiap bagian dan memuat data penting dari masing-masing bagian tersebut. Terdapat beberapa jenis peringkasan, di antaranya:

a. Ekstraksi berbasis (extraction-based summarization)

Metode ekstraksi melibatkan pemilihan unit teks (kalimat, segmen kalimat, paragraf, atau bagian) yang dianggap memiliki informasi relevan, dan kemudian mengatur unit tersebut dengan tepat. Terdapat tiga kategori algoritma berbasis ekstraksi:

1. Surface-level

Metode penilaian kalimat (sentence scoring) menggunakan berbagai fitur seperti TF/IDF, huruf kapital, kata benda, frasa tanda, data numerik, panjang kalimat, letak kalimat, dan kemiripan judul. Skor-skor ini kemudian diolah menggunakan data pelatihan untuk memutuskan apakah sebuah kalimat masuk ke dalam ringkasan atau tidak.

2. Intermediate-level

Metode peringkasan ekstraktif dengan menggunakan rantai leksikal (lexical chain) telah diteliti oleh Barzilay (1997). Metode ini menggunakan rantai leksikal sebagai model subjek untuk peringkasan.

3. Deep parsing

Rhetorical Structure Theory (RST) adalah kerangka kerja yang digunakan untuk mengidentifikasi struktur teks pada tingkat klausa. Chengcheng (2010) menggabungkan RST dengan metode peringkasan

untuk memisahkan bagian yang kurang penting dan mengekstrak teks struktur retorikal berdasarkan relasi retorikal antara kalimat.

b. Abstraksi (abstractive summarization)

Teknik abstraksi memanfaatkan paraphrasing section dari sumber dokumen untuk menghasilkan ringkasan. Berbeda dengan ekstraksi, abstraksi dapat menghasilkan ringkasan yang lebih kuat karena dapat mengungkapkan informasi yang tidak secara eksplisit disajikan dalam teks asli. Namun, pengembangan program untuk teknik ini lebih sulit karena melibatkan penggunaan teknologi natural language generation yang masih terus berkembang. Evaluasi hasil rangkuman dapat diklasifikasikan menjadi dua jenis, yaitu evaluasi intrinsik dan ekstrinsik. Evaluasi intrinsik dilakukan dengan membandingkan kualitas rangkuman dengan standar ideal. Sementara itu, evaluasi ekstrinsik mengukur sejauh mana rangkuman membantu kinerja dalam tugas-tugas tertentu.

2.2 *Algoritma TextRank*

TextRank adalah teknik peringkat berbasis grafik yang mengekstraksi grafik dari teks menggunakan teks bahasa alami. Grafik *TextRank* adalah grafik berbobot tidak terarah. Sebuah simpul penting dalam sebuah graf dapat diidentifikasi menggunakan algoritma pemeringkatan berbasis graf, yang menggunakan informasi global yang dideskripsikan secara rekursif dari seluruh graf. Konsep dasarnya adalah mengintegrasikan voting atau saran dari node grafik (Zamzam, 2020). Node a benar-benar memberikan suara untuk node b ketika mereka ditautkan. Kepentingan sebuah simpul ditentukan oleh skornya. Sehingga semakin besar skornya, semakin signifikan node tersebut. Oleh karena itu, skor yang diberikan ke sebuah node ditetapkan berdasarkan suara yang diterima node dari node lain serta skor node yang menerima suara.

Sebagai ilustrasi, graf G = (V,E) E adalah bagian dari VxV, dan V adalah himpunan simpul dan sisi. Rumus kesamaan dapat menghitung bobot sisi, yang memiliki nilai. Ada dua set simpul pada simpul Vi: In (Vi), yang merupakan simpul yang menunjuk ke Vi, dan Keluar (Vi), yang merupakan simpul yang menunjuk ke Vi (penerus). Seperti yang didefinisikan di bawah ini, skor keseluruhan pada simpul Vi:

$$S(Vi) = (1-d) + d * \sum_{i \epsilon In(Vi)} \frac{1}{|Out(Vi)|} S(Vi) \quad (1)$$

Dimana d adalah faktor redaman, yang berkisar dari 0 hingga 1. Probabilitas perjalanan dari satu node acak ke node acak lainnya pada grafik diintegrasikan berdasarkan nilai d. Algoritma ini mengimplementasikan "random surfer model" dalam konteks penjelajahan web, di mana pengguna secara acak mengklik tautan dengan probabilitas d dan menavigasi ke situs web yang sama sekali berbeda dengan probabilitas 1-d. Faktor d biasanya 0,85.

Nilai simpul baru diperoleh dengan setiap iterasi. Metode berikut dapat digunakan untuk menentukan kesalahan dengan membandingkan hasil dari iterasi terbaru dengan iterasi terbaru:

$$S^{k+1}(Vi) - S^{k}(Vi) \quad (2)$$

Setelah metode dijalankan, skor ditautkan pada setiap simpul, yang menunjukkan "kepentingan" simpul dalam grafik. Nilai awal yang dipilih tidak berdampak pada nilai akhir setelah *TextRank* selesai, namun berdampak pada jumlah iterasi yang diperlukan untuk mencapai ambang batas.

Karena pendekatan yang digunakan menggunakan graf berbobot, penting untuk memahami seberapa besar kontribusi masing-masing sisi terhadap hasil akhir. Rumus *TextRank* adalah sebagai berikut:

$$W\,S(Vi) = (1-d)\; + d * \sum_{Vi \epsilon In(Vi)} \frac{\omega ji}{\sum_{Vk \epsilon Out(vi)} \omega jk}\, W\,S(Vi) \quad (3)$$

Rincian Rumus:

WS(Vi)	= Bobot simpul
Vi & d	= damping factor (0.85)
Vk	= Himpunan simpul tetangga
Vj WS(Vi)	= Bobot simpul
Vj Vj	= Himpunan simpul tetangga
Vi wji & wjk	= bobot sisi antara simpul j&i dan j&k secara berurutan.

2.3 Proses *Algoritma TextRank*

TextRank adalah algoritma yang digunakan untuk mengekstrak ringkasan dari teks. Algoritma ini didasarkan pada *PageRank*, algoritma yang digunakan oleh mesin pencari *Google* untuk menentukan relevansi halaman web.

Langkah-langkah cara kerja algoritma *TextRank* adalah sebagai berikut:

a. *Preprocessing*

Tahap ini melibatkan pembersihan dan pra-pengolahan teks seperti menghapus tanda baca, mengubah teks ke dalam huruf kecil, dan memisahkan kata menjadi frasa.

b. Tokenisasi

Tahap ini melibatkan pemecahan teks menjadi token atau kata individu. Tokenisasi pada *TextRank* adalah proses membagi teks menjadi unit-unit yang lebih kecil yang disebut "*token*" atau "kata-kata". Tokenisasi adalah langkah penting dalam pemrosesan bahasa alami yang melibatkan algoritma seperti *TextRank*, yang digunakan untuk analisis teks dan penggalian informasi. Dalam konteks *TextRank*, *tokenisasi* umumnya melibatkan pemisahan kalimat menjadi kata-kata individu. Pemisahan ini biasanya dilakukan dengan menggunakan aturan-aturan linguistik dan metode pemrosesan teks yang mengidentifikasi titik akhir kalimat, spasi antara kata-kata, dan tanda baca lainnya. Setelah teks telah di-*tokenisasi, token-token* ini menjadi unit-unit yang dapat diproses lebih lanjut oleh algoritma *TextRank*. Algoritma ini membangun graf berbobot di mana setiap token mewakili simpul dan hubungan antara token-token tersebut ditentukan berdasarkan kedekatan dan hubungan semantik di antara mereka.

Graf ini kemudian digunakan untuk menghitung skor pentingnya setiap *token* dalam teks. Dengan kata lain, tokenisasi pada *TextRank* adalah langkah awal dalam memproses teks menggunakan algoritma tersebut. Ini memungkinkan analisis lebih lanjut terhadap teks dan identifikasi *token-token* yang paling penting atau relevan dalam konteks yang diberikan.

c. Representasi Vektor

Setiap token dalam teks diwakili oleh vektor representasi yang menyimpan informasi tentang konteks dan frekuensi token. Reprentasi vektor *TextRank* merupakan salah satu teknik pemrosesan bahasa alami yang digunakan untuk menganalisis dan memahami teks secara otomatis. Teknik ini berfokus pada pendekatan berbasis graf untuk menentukan pentingnya suatu kata atau frasa dalam sebuah dokumen teks.
Pertama, dokumen teks dipecah menjadi unit-unit yang lebih kecil, seperti kalimat atau frasa. Selanjutnya, dibangunlah graf berbentuk jaringan di mana setiap unit teks menjadi simpul (node) dalam graf tersebut. Keterhubungan antar unit teks diwakili oleh hubungan antara simpul-simpul dalam graf. Setelah itu, algoritma *TextRank* digunakan untuk menghitung skor pentingnya setiap simpul (unit teks) dalam graf. Skor ini dihitung berdasarkan faktor-faktor seperti jumlah hubungan yang dimiliki oleh simpul tersebut dan skor pentingnya simpul-simpul terkait yang terhubung dengannya. Proses ini mencoba merepresentasikan tingkat kepentingan suatu kata atau frasa dalam konteks dokumen secara keseluruhan. Akhirnya, hasil dari algoritma *TextRank* adalah representasi vektor yang menggambarkan tingkat kepentingan setiap unit teks dalam dokumen. Representasi vektor ini dapat digunakan untuk berbagai tujuan, seperti analisis teks, ekstraksi informasi, atau pemeringkatan dokumen berdasarkan tingkat kepentingannya. Dengan menggunakan representasi vektor *TextRank,* kita dapat mengidentifikasi kata atau frasa yang paling penting dalam suatu teks. Hal ini memungkinkan pengambilan keputusan yang lebih baik berdasarkan informasi yang dihasilkan, serta membantu dalam pemahaman dan analisis teks secara efisien.

d. *Similarity Matrix (Cosine Similarity)*

Tahap ini melibatkan pembentukan matriks keterkaitan antara frasa yang berdekatan dalam teks. *Cosine similarity* adalah metrik yang umum digunakan dalam *TextRank* untuk mengukur kesamaan antara dua vektor representasi teks. Pada *TextRank, cosine similarity* digunakan untuk mengukur kesamaan antara vektor representasi token-token dalam sebuah dokumen. Langkah-langkah umum untuk menghitung *cosine similarity* dalam *TextRank* adalah sebagai berikut: *Tokenisasi* dan Preprocessing: Dokumen atau teks yang akan diproses di-*tokenisasi* menjadi unit-unit yang lebih kecil seperti kata-kata. Selain itu, langkah-langkah *preprocessing* seperti menghilangkan kata-kata yang tidak relevan (*stop words*), stemming, atau lematization juga dapat dilakukan untuk memperbaiki representasi teks. Pembentukan Matriks Bobot: Matriks bobot dibangun berdasarkan hubungan antara token-token dalam dokumen. Matriks ini

dapat berupa *matriks adjacency* atau matriks *term frequency-inverse document frequency (TF-IDF)* yang merepresentasikan hubungan dan bobot antara token-token tersebut. Perhitungan *Cosine Similarity*: Setelah matriks bobot terbentuk, *cosine similarity* dihitung untuk setiap pasangan token dalam dokumen. *Cosine similarity* mengukur sudut antara dua vektor representasi token, dan semakin kecil sudutnya, semakin mirip token-token tersebut. Pembentukan Graf dan Perangkingan: Berdasarkan nilai *cosine similarity*, sebuah graf berbobot dibentuk di mana setiap token mewakili simpul dan bobot edge antara token-token ditentukan oleh *cosine similarity* mereka. Graf ini kemudian digunakan untuk melakukan algoritma *TextRank* yang menghasilkan peringkat *token-token* berdasarkan pentingnya. Dengan menggunakan *cosine similarity, TextRank* dapat mengidentifikasi token-token yang memiliki hubungan dan kemiripan semantik yang tinggi dalam teks. Ini memungkinkan untuk menghasilkan peringkat token-token yang paling penting dan relevan dalam konteks yang diberikan.

e. *Scoring*

Setiap frasa dalam teks diberikan skor berdasarkan hubungan antar frasa dalam matriks keterkaitan. *Scoring* dalam algoritma *TextRank* mengacu pada perhitungan skor pentingnya token-token dalam teks. Skor ini digunakan untuk merangkingkan *token-token* dan mengidentifikasi *token-token* yang paling penting atau relevan dalam teks tersebut. Berikut adalah langkah-langkah umum dalam proses *scoring* menggunakan *TextRank: Tokenisasi* dan *Preprocessing*: Dokumen atau teks di-*tokenisasi* menjadi *token-token* yang lebih kecil, seperti kata-kata. Langkah-langkah preprocessing seperti penghilangan *stop words and stemming* juga dapat dilakukan untuk meningkatkan representasi teks. Pembentukan Graf Berbobot: Berdasarkan *token-token* yang dihasilkan, sebuah graf berbobot dibentuk. Setiap *token* direpresentasikan sebagai simpul dalam graf, dan hubungan antara *token-token* ditentukan oleh kedekatan dan hubungan semantik di antara mereka. Graf ini dapat direpresentasikan dalam bentuk matriks *adjacency* atau matriks bobot lainnya. Iterasi Algoritma *TextRank*: Algoritma *TextRank* diiterasikan hingga *konvergens*i atau jumlah iterasi yang ditentukan sebelumnya. Dalam setiap iterasi, skor pentingnya setiap token diperbarui berdasarkan skor token-token yang terhubung dengannya dalam graf. Skor dapat dihitung menggunakan metode seperti *PageRank,* yang mengukur pentingnya simpul berdasarkan jumlah dan pentingnya simpul-simpul yang terhubung dengannya. Perankingan *Token:* Setelah algoritma *TextRank* mencapai *konvergensi, token-token* diberi skor penting berdasarkan hasil akhir perhitungan. *Token-token* dengan skor tertinggi dianggap sebagai *token-token* yang paling penting atau relevan dalam teks. Hasil ini dapat diurutkan dalam urutan penurunan skor untuk menentukan peringkat *token-token*. Melalui proses *scoring* ini, *TextRank* dapat menghasilkan peringkat token-token yang dapat digunakan untuk memahami struktur dan isi teks, serta mengidentifikasi token-token yang paling penting atau relevan dalam analisis teks.

f. Rangking

Tahap ini melibatkan pengurutan frasa dalam teks berdasarkan skor yang diberikan. Ranking adalah langkah dalam *TextRank* yang melibatkan pengurutan *token* berdasarkan skor penting mereka. Setelah setiap *token* diberi skor*, token-token* tersebut diurutkan secara menurun berdasarkan skor tersebut. Hasilnya adalah daftar *token* yang diberi peringkat berdasarkan kepentingan relatif mereka dalam teks. ranking (perankingan) adalah pengurutan *token* berdasarkan skor yang dihasilkan dari proses *scoring. Scoring* memberikan nilai numerik yang menggambarkan pentingnya setiap token, sedangkan ranking menempatkan token-token tersebut dalam urutan berdasarkan skor tersebut

g. Ekstraksi

Frasa dengan skor tertinggi dipilih sebagai bagian teks yang akan disertakan dalam ringkasan. Ekstraksi menggunakan algoritma *TextRank* adalah proses untuk mengidentifikasi dan mengekstraksi informasi yang penting dari teks berdasarkan peringkat token-token dalam teks tersebut.

Proses algoritma *TextRank* berlangsung secara iteratif sampai skor frasa tidak lagi berubah atau mencapai kondisi *konvergensi*. Hasil akhir adalah ringkasan teks yang mencakup frasa yang paling penting dan relevan.

2.4 *ROUGE*

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) adalah sekumpulan metrik dan perhitungan komputer yang umumnya digunakan untuk mengevaluasi peringkasan teks otomatis dan terjemahan otomatis yang dihasilkan oleh komputer. *ROUGE* sangat penting dalam bidang ilmu pemrosesan bahasa alami (*natural language processing*). Algoritma ini digunakan untuk menilai kualitas peringkasan teks otomatis yang dibuat oleh sistem dengan membandingkannya dengan ringkasan ideal yang dibuat oleh manusia. Langkah-langkah dalam algoritma ini mencakup menghitung unit-unit yang sama dan membandingkannya dalam beberapa aspek penilaian, seperti n-gram, urutan kata, dan kesamaan antara unit-unit. Hasil evaluasi dilakukan langsung oleh manusia yang melakukan peringkasan (Riyadi & Samiati, 2021).

a. *ROUGE-N*

ROUGE-N adalah metode perhitungan *recall* yang didasarkan pada perbandingan n-gram antara ringkasan standar (*gold standard summary*) dan teks hasil peringkasan mesin. Penggunaan n-gram dapat bervariasi, dari n=1 hingga n=4. Namun, n-gram yang paling umum digunakan adalah n=1 (*ROUGE*-1) dan n=2 (*ROUGE*-2). Misalnya, p merupakan jumlah n-gram yang sama antara ringkasan standar dan teks hasil peringkasan mesin, sedangkan q merupakan jumlah n-gram pada ringkasan standar (Yuliska & Syaliman, 2020). Berikut rumus perhitungan ROUGE-N:

$$ROUGE - N = \frac{P}{q} \tag{4}$$

b. *ROUGE-L*

ROUGE-L digunakan untuk mengevaluasi tingkat kesamaan antara ringkasan teks mesin dan ringkasan standar (*gold standard summary*) dengan membandingkan "*longest common subsequence*" (LCS) atau rangkaian kata

terpanjang yang sama di antara keduanya. Jumlah kata pada ringkasan standar dapat direpresentasikan dengan variabel m (Rojas-Simón et al., 2021). Berikut rumus perhitungan ROUGE-L:

1) *Precision*

$$\boldsymbol{P_{lcs}} = \frac{\boldsymbol{LCS(x,y)}}{\boldsymbol{n}} \quad (5)$$

2) *Recall*

$$\boldsymbol{R_{lcs}} = \frac{\boldsymbol{LCS(x,y)}}{\boldsymbol{m}} \quad (6)$$

3) *F1 Score*

$$\boldsymbol{F_{lcs}} = \frac{\boldsymbol{2 * (P * R)}}{\boldsymbol{(P+R)}} \quad (7)$$

Keterangan:
P = *Precision*
R = *Recall*
F = *F1 Score*
x , y = Jumlah kesamaan antar kata ringkasan dan teks asli.
n = Jumlah total unigram dalam ringkasan.
m = Jumlah total bigram dalam referensi.

Penjelasan Singkat *ROUGE* adalah sebuah sekumpulan metrik dan perhitungan computer dan mengukur sejauh mana system dapat menghasilakn rinskasan informasi yang penting dari teks referensi.
Dalam percarian nilai akurat pada *ROUGE* sama *menggunakan Precision, Recall,* dan *F1 Score.*

Precision untuk mengukur berapa banyak kata dan frasa yang sama dalam teks referensi dan ringkasan.

Recall untuk mengukur tingkat kelengkapan dan kemampuan system dalam menangkap informasi penting dalam teks referensi.

F1 Score untuk menghitung rata-rata dari *precision* dan *recall* untuk mengukur poerforma sistem. Jika Performa tinggi yang dihasilkan maka sistem tersebut baik dalam menghasilkan teks ringkasan.