PERINGKASAN TEKS OTOMATIS (*AUTOMATED TEXT SUMMARIZATION*) PADA ARTIKEL BERBAHASA INDONESIA MENGGUNAKAN ALGORITMA *LEXRANK*

TUGAS AKHIR

Disusun Sebagai Salah Satu Syarat
Untuk Memperoleh Gelar Sarjana Teknik
Pada Jurusan Teknik Informatika

Oleh

HALIMAH

NIM. 11850124454

FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI SULTAN SYARIF KASIM RIAU
PEKANBARU
2023

LEMBAR PERSETUJUAN

PERINGKASAN TEKS OTOMATIS (*AUTOMATED TEXT SUMMARIZATION*) PADA ARTIKEL BERBAHASA INDONESIA MENGGUNAKAN ALGORITMA *LEXRANK*

TUGAS AKHIR

Oleh

HALIMAH
NIM. 11850124454

Telah diperiksa dan disetujui sebagai Laporan Tugas Akhir
di Pekanbaru, pada tanggal 12 Januari 2023

Pembimbing I,	Pembimbing II,
SURYA AGUSTIAN, S.T., M.KOM.	**SITI RAMADHANI, S.PD,M.KOM.**
NIP. 19760830 201101 1 003	**NIK. 130 517 045**

LEMBAR PENGESAHAN

PERINGKASAN TEKS OTOMATIS (*AUTOMATED TEXT SUMMARIZATION*) PADA ARTIKEL BERBAHASA INDONESIA MENGGUNAKAN ALGORITMA *LEXRANK*

Oleh

HALIMAH

NIM. 11850124454

Telah dipertahankan di depan sidang dewan penguji
sebagai salah satu syarat untuk memperoleh gelar Sarjana Teknik
pada Universitas Islam Negeri Sultan Syarif Kasim Riau

Pekanbaru, 12 Januari 2023

Mengesahkan,

Dekan,	Ketua Jurusan,
DR. HARTONO, M.PD.	**IWAN ISKANDAR, M.T.**
NIP. 19640301 199203 1 003	**NIP. 19821216 201503 1 003**

DEWAN PENGUJI

Ketua	: Iwan Iskandar, M.T.
Pembimbing I	: Surya Agustian, S.T, M.Kom.
Pembimbing II	: Siti Ramadhani, S.Pd, M.Kom.
Penguji I	: Muhammad Fikry, ST, M.Sc.
Penguji II	: Febi Yanto, M.Kom.

SURAT PERNYATAAN

Saya yang bertanda tangan dibawah ini :

Nama : Halimah

NIM : 11850124454

Tempat/Tgl.Lahir : Rantau Bais/08-Mei-1999

Fakultas/Pascasarjana : Sains dan Teknologi/S1

Prodi : Teknik Informatika

Judul Skripsi :

PERINGKASAN TEKS OTOMATIS (*AUTOMATED TEXT SUMMARIZATION*) PADA ARTIKEL BERBAHASA INDONESIA MENGGUNAKAN ALGORITMA *LEXRANK*

Menyatakan dengan sebenar-benarnya bahwa :

1. Penulisan Skripsi dengan judul sebagaimana tersebut di atas adalah hasil pemikiran dan penelitian saya sendiri.
2. Semua kutipan pada karya tulis saya ini sudah disebutkan sumbernya.
3. Oleh karena itu Skripsi saya ini,saya nyatakan bebas dari plagiat.
4. Apabila dikemudian hari terbukti terdapat plagiat dalam penulisan Skripsi saya tersebut, maka saya bersedia menerima sanksi sesuai peraturan perundang-undangan.

Demikian surat pernyataan ini saya buat dengan penuh kesadaran tanpa paksaan dari pihak manapun juga.

Pekanbaru, 12 Januari 2023

Yang membuat pernyataan

Halimah

NIM. 11850124454

pilih salah satu sesuai jenis karya tulis

LEMBAR HAK ATAS KEKAYAAN INTELEKTUAL

Tugas Akhir yang tidak diterbitkan ini terdaftar dan tersedia di Perpustakaan Universitas Islam Negeri Sultan Syarif Kasim Riau adalah terbuka untuk umum dengan ketentuan bahwa hak cipta pada penulis. Referensi kepustakaan diperkenankan dicatat, tetapi pengutipan atau ringkasan hanya dapat dilakukan seizin penulis dan harus disertai dengan kebiasaan ilmiah untuk menyebutkan sumbernya.

Penggandaan atau penerbitan sebagian atau seluruh Tugas Akhir ini harus memperoleh izin dari Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Sultan Syarif Kasim Riau. Perpustakaan yang meminjamkan Tugas Akhir ini untuk anggotanya diharapkan untuk mengisi nama, tanda peminjaman dan tanggal pinjam.

LEMBAR PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Tugas Akhir ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan didalam daftar pustaka.

Pekanbaru, 12 Januari 2023

Yang membuat pernyataan,

HALIMAH

NIM. 11850124454

LEMBAR PERSEMBAHAN

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

Dengan menyebut nama Allah Yang Maha Pengasih lagi Maha Penyayang Alhamdulillahirobbil'alamiin, puji syukur kepada Allah SWT karena berkat rahmat dan karunianya sehingga saya dapat menyeselesaikan tugas akhir ini. Shalawat beserta salam tak lupa saya hadiahkan kepada baginda Nabi Muhammad SAW, dengan mengucapkan Allahumma Sholli'ala Muhammad wa'alaali Muhammad.

Ayah dan ibu tercinta

Karya sederhana ini Halimah persembahan untuk kedua orangtuaku tercinta. "Maaf ayah….. putri kecil mu ini dulu belum bisa membahagiankanmu, sedih sekali rasanya disaat putrimu ini berhasil sampai di titik ini tapi ayah sudah tidak ada di sampingku, aku sayang banget sama ayah tapi sayangnya Allah jauh lebih sayang sama ayah, bahagia yaa di surga sana do'a ku tiada putus untukmu ayah tersayang. Love you ayah….."

"Ibu terimakasih untuk selama ini sudah sangat sabar menunggu putri kecilmu ini hingga sampai di titik ini. Ibu, terimakasih untuk setiap pelukanmu, segala nasehat-nasehatmu, dukungan dan setiap do'a yang menyertaiku, mungkin aku tidak akan bisa membalas semuanya jasamu ibu. Love you ibu….."

Diriku sendiri

"Terimakasih untuk diriku sendiri sudah bertahan sampai di titik ini, kamu hebat. Banyak suka-duka yang menghampiri untuk bisa sampai ke titik ini, pernah ingin menyerah tapi senyum dan harapan orangtualah yang menguatkanku untuk sampai ketitik ini."

Dosen Pembimbing Tugas Akhir

"Bapak Surya Agustian,S.T.,M.Kom., dan Ibu Siti Ramadhani, S.Pd,M.Kom., selaku dosen pembimbing Tugas Akhir. Terima kasih banyak kepada Bapak/Ibu sudah membantu Halimah selama ini, memberikan nasehet, bimbingan, serta arahan sampai Tugas Akhir ini selesai".

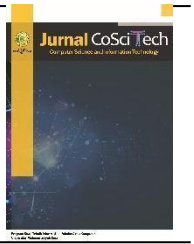
Vol. 3, No. 3, Desember 2022, hal. 371-381

CoSciTech
Computer Science and Information Technology

p-ISSN: 2723-567X
e-ISSN: 2723-5661

Jurnal Computer Science and Information Technology (CoSciTech)

http://ejurnal.umri.ac.id/index.php/coscitech/index



Peringkasan teks otomatis (automated text summarization) pada artikel berbahasa indonesia menggunakan algoritma lexrank

Hakman[*], Surya Agustian[2], Siti Ramadhani[3]

Email: [1]11850124454@students.uin-suska.ac.id, [2]surya.agustian@uin-suska.ac.id , [3]siti.ramadhani@uin-suska.ac.id

[1,2,3]Teknik Informatika, Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau

Diterima: 15 November 2022 | Direvisi: - | Disetujui: 16 Desember 2022

Abstrak

Artikel merupakan kumpulan teks atau kalimat yang panjang dan berisikan gagasan atau pendapat terhadap suatu topik tertentu. Artikel yang sangat panjang akan menghabiskan waktu cukup lama untuk membaca dan memahami poin-poin penting yang disampaikan. Penelitian ini mengusulkan algoritma *Lexrank* untuk meringkas teks otomatis pada artikel berbahasa indonesia. Penelitian ini menggunakan dataset berupa korpus yang tersusun atas 300 artikel dari berbagai topik. Kalimat yang dipilih menjadi ringkasan untuk *gold standard* dirata-ratakan dari dua orang *annotator*. Metode peringkas dokumen dikembangkan untuk menghasilkan ringkasan yang performanya dibandingkan dengan *gold standard* tersebut menggunakan *ROUGE score*. Metode bekerja dengan beberapa tahap, mulai dari *text preprocessing* yang meliputi segmentasi kalimat, *case folding*, tokenisasi, *punctuation removal, stemming* dan *stopword removal*. Kemudian menghitung bobot tf-idf, bobot *similarity*, pembentukan *graf*, pemeringkatan kalimat, dan tahap akhir adalah membentuk hasil ringkasan. Untuk pengembangan sistem, 150 dokumen diuji coba dengan variasi pemilihan *ranking similarity*, dan yang terbaik digunakan sebagai model untuk *test document*. Hasil pengujian dengan *compression rate* 50% menghasilkan nilai *f-measure* rata-rata untuk 150 *test document*, pada metrik *ROUGE-1, ROUGE-2* dan *ROUGE-L* secara berturut-turut adalah 67,53%, 59,10%, dan 67,05%. Sedangkan untuk *compression rate* 30% menghasilkan rata-rata *f-measure* pada *ROUGE-1, ROUGE-2* dan *ROUGE-L* secara berturut-turut adalah 55,82%, 45,51%, dan 54,76%. Penelitian ini menghasilkan akurasi *f-measure* yang lebih baik dan kompetitif bila merujuk pada hasil-hasil penelitian sejenis.

Kata kunci: *ringkasan, rouge, sistem peringkasan otomatis , lexrank*

Automatic text summarization in indonesian articles using the lexrank

Abstract

An article is a collection of text or sentences that are long and contain ideas or opinions on a particular topic. A very lon g article will take a long time to read and understand the key points presented. This study proposes a Lexrank algorithm for summarizing automatic text in Indonesian-language articles. This study used a dataset in the form of a corpus composed of 300 articles from various topics. The sentence chosen to be a summary for the gold standard is averaged from two annotators. The document summarization method was developed to produce a summary whose performance was compared to the gold standard using the ROUGE score. The method works with several stages, ranging from text preprocessing which includes sentence segmentation, case folding, tokenization, punctuation removal, stemming and stopword removal. Then calculate the weight of tf.idf, the weight of similarity, the formation of the graph, the ranking of sentences, and the final stage is to form a summary result. For system development, 150 documents were tested with variations in ranking similarity selection, and the best ones were used as models for test documents. The test results with a compression rate of 50% resulted in average f-measure values for 150 test documents, on the ROUGE-1, ROUGE-2 and ROUGE-L metrics respectively were 67.53%, 59.10%, and 67.05%. As for the 30% compression rate, the average f-measure on ROUGE-1, ROUGE-2 and ROUGE-L respectively is 55.82%, 45.51%, and 54.76%. This study resulted in better and more competitive f-measure accuracy when referring to similar research results.

Keywor ds: *summarization, rouge, automatic text summarization , lexrank*

doi: https://doi.org/10.37859/coscitech.v3i3.4300

1. PENDAHULUAN

Teknologi informasi yang berkembang semakin cepat menyebabkan penyebaran informasi dapat diakses secara bebas melalui internet. Beragamnya informasi seperti berita, artikel, iklan, promosi, dan pengumuman, disebarkan secara online. Kemudahan akan informasi ini menimbulkan peluang tersampainya suatu informasi ke tujuan menjadi lebih besar. Informasi yang banyak diakses secara mudah dari internet dapat berbentuk artikel, makalah, laporan, liputan, dan sebagainya. Artikel merupakan tulisan yang terdiri dari teks panjang berisikan pikiran, pendapat, hingga kritik terhadap suatu persoalan yang sedang berkembang di masyarakat. Namun, artikel yang sangat panjang akan menghabiskan waktu cukup lama untuk memahami poin-poin utama artikel, maka diperlukan sistem peringkasan teks otomatis untuk menyelesaikan ekstraksi informasi penting dalam sebuah artikel [1]. Salah satu studi yang bisa membantu menyelesaikan masalah ini adalah peringkasan teks (*automated text summarization*).

Peringkasan teks otomatis (*automated text summarization*) merupakan tahapan untuk mendapatkan informasi dari sebuah dokumen atau teks menggunakan komputer. Dengan adanya sistem ini dapat membantu pembaca memahami suatu artikel tanpa menghilangkan informasi, sehingga pembaca tidak perlu membaca seluruh isi dokumen, pembaca dapat memahami dan mengetahui informasi penting dari dokumen serta menghemat waktu pembaca. Peringkasan teks otomatis termasuk dalam studi pemprosesan bahasa alami (*natural language processing*). Peringkasan teks otomatis akan menghasilkan dokumen yang biasanya tidak melebihi dari 50% ukuran dokumen sumber. Dengan jumlah kalimat yang diekstraksi adalah 25%-50% dari total kalimat dalam dokumen [2]. Ada dua tahap peringkasan teks otomatis yaitu, peringkasan ekstraktif (*extractive summarization*) dilakukan proses pengambil sejumlah kalimat penting dari dokumen teks aslinya tanpa mengubah kalimatnya. Peringkasan abstraktif (*abstractive summarization*) dilakukan dengan cara merubah dan merangkai kalimat baru yang sebenarnya intisari dari informasi pada dokumen yang diringkas [3]. Untuk tahap pengujian peringkasan teks otomatis akan menggunakan *ROUGE. ROUGE* merupakan standar pengujian peringkasan teks berbahasa indonesia untuk mengukur kualitas hasil ringkasan dengan membandingkan hasil peringkasan yang dilakukan oleh sistem dan hasil ringkasan oleh manusia(manual) dengan mencari nilai dari *precision, recall,* dan *f-measure*.

Metode peringkasan teks otomatis berdasarkan pemeringkatan berbasis *graf*, antara lain adalah algoritma *lexrank*. Algoritma ini dapat digunakan untuk meringkas dokumen tunggal atau multidokumen. *Lexrank* dengan pendekatan *centroid* dalam [4] sukses dalam meringkas banyak dokumen dibandingkan dengan sistem peringkasan lainnya. *Lexrank* diperlukan pada sistem peringkas dokumen yang lebih besar untuk menyatukan antara bobot dan hal-hal lain seperti panjang, letak kalimat, dengan menggunakan gabungan linear yang bobotnya bisa diatur oleh pengguna.

Berdasarkan penelitian oleh [5] algoritma *Lexrank* mampu memperbaiki panjang ringkasan dan memastikan bahwa tidak ada poin penting yang terlewatkan dalam teks ringkasan. Penelitian yang menggabungkan teknik *ekstraksi* dan metode *Cross Latent Semantic Analysis (CLSA)* oleh [6] menghasilkan rata-rata nilai akurasi *f-measure, precision,* dan *recall* pada *compression rate* 20% berturut-turut yaitu 0.3853, 0.432, dan 0.3715. Penelitian ini menghasilkan akurasi cukup rendah dibandingkan penelitian sebelumnnya yang menggunakan algoritma yang sama.

Metode lainnya untuk peringkasan teks otomatis menggunakan *TextRank* , misalnya pada [7] menghasilkan rata-rata nilai *F-Score* 0.439 pada *ROUGE*-1 dan 0,3186 pada *ROUGE*-2. Sedangkan *TextRank* yang dimodifikasi mendapatkan rata-rata nilai *F Score* 0,3999 pada *ROUGE*-1 dan 0,2805 pada *ROUGE*-2, hasilnya tidak jauh berbeda dan masih tergolong rendah.

Penerapan teks mining dan *lexrank* pada [2] mempermudah dan mempercepat proses pencarian yang dilakukan oleh pengguna dalam mencari sebuah teks. Sedangkan pada [8], metode *lexrank* yang telah dimodifikasi dipakai untuk meringkas *tweet* berdasarkan waktu *tweet* dari aspek temporal dari 400 juta *tweet*. Metode ini dapat membantu seseorang dengan mudah menghasilkan ringkasan *tweet* dan dapat memperoleh rincian hingga waktu tertentu sehingga meningkatkan metode peringkasan.

Penelitian ini menerapkan algoritma *lexrank* pada peringkasan teks otomatis pada artikel berbahasa Indonesia, dan menyelidiki pengaruh pemilihan kandidat kalimat dari ranking yang dihitung oleh *lexrank* terhadap hasil ringkasan. Kualitas hasil peringkasan diperiksa menggunakan *ROUGE score* berdasarkan *gold standard* ringkasan yang disusun oleh manusia (*human annotator*). Bagian selanjutnya dari paper ini menjelaskan metode peringkasan dokumen yang diusulkan, dan rancangan variasi yang dilakukan untuk eksperimen. Beberapa temuan menarik dilaporkan di dalam bagian hasil dan analisa, dan selanjutnya ditutup dengan bagian kesimpulan.

2. METODE PENELITIAN

2.1. Penyusunan Dataset

Penelitian ini dimulai dari pengumpulan 300 artikel berita dari internet. Sumber artikel dari berbagai topik seperti gaya hidup, hiburan, kesehatan, politik, teknologi, kriminal, dan lain-lain, dengan jumlah sebagaimana diuraikan pada Tabel 1. Dari setiap artikel yang diambil, kemudian diberikan *score* oleh 2 orang *annotator*. *Score* diberikan untuk setiap kalimat di dalam artikel, seberapa penting kalimat tersebut untuk dipilih sebagai ringkasan. *Score* setiap kalimat kemudian dirata-ratakan, dan diranking

doi: https://doi.org/10.37859/coscitech.v3i3.4300

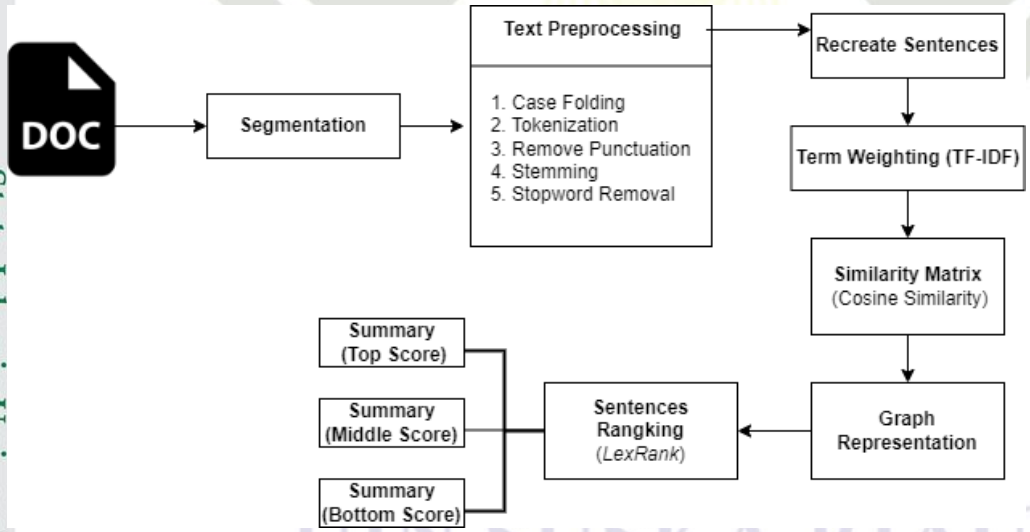
dari *score* yang paling tinggi. Kalimat yang terpilih sebagai ringkasan, adalah sejumlah tertentu berdasarkan tingkat kompresi (*compression rate*) yang ditentukan pada eksperimen.

Tabel 1. Data Statistik dari Topik Artikel pada Korpus Penelitian

No	Topik	Jumlah artikel
1	Agama	44
2	Budaya	10
3	Ekonomi	19
4	Gaya hidup	11
5	Hiburan	20
6	Kriminal	20
7	Kesehatan	34
8	Lingkungan	15
9	News	23
10	Olahraga	11
11	Pendidikan	13
12	Politik	29
13	Sosial	15
14	Teknologi	36
	Total	**300**

2.2. Algoritma *LexRank*

Algoritma *lexrank* menghasilkan ringkasan kalimat kalimat menjadi sebuah ringkasan[9]. Berikut merupakan alur algoritma *lexrank* menghasilkan sebuah ringkasan. Ringkasan yang dibangkitkan terdiri dari 3 jenis, yaitu ringkasan yang diambil dari kalimat-kalimat dengan *sentence ranking* berada pada *top-n*, *bottom-n* dan *middle-n score*, dengan n adalah jumlah kalimat sesuai dengan tingkat kompresi yang diinginkan. Dalam penelitian ini, diselidiki tingkat kompresi sebesar 50% dan 30%.



Gambar 1. Diagram sistem peringkas dokumen dalam penelitian ini

2.2.1. Pemecahan Kalimat(*Segmentation*)

Pada tahap segmentasi, dokumen akan dipecah menjadi beberapa kalimat berdasarkan tanda pemisah seperti tanda titik ".", tanda tanya "?" dan tanda seru "!". Setiap dokumen yang telah dipecah akan dimasukkan ke dalam list kalimat.

2.2.2. Prapemrosesan Teks(*Text Preprocessing*)

Tahap *Preprocessing* merupakan langkah awal untuk membuat pemrosesan teks menjadi lebih terstruktur [10]. Tujuan pemrosesan ini untuk mempersiapkan teks menjadi data yang siap diproses. Berikut merupakan tahapan-tahapan dalam *text preprocessing* sebagai berikut:

doi: https://doi.org/10.37859/coscitech.v3i3.4300

1. *Case Folding*

Tahap *case folding* mengubah semua kata menjadi huruf kecil dan menghapus tanda baca selain 'a-z', angka, dan tanda baca yang dianggap tidak perlu.

2. *Tokenizing*

Pada tahap *tokenizing*, kalimat hasil *case folding* dipecah menjadi perkata. Pemecahan kalimat kedalam kata berdasarkan tanda spasi antar kalimat, kemudian dibuatlah list yang terdiri dari kumpulan kata yang disebut token.

3. *Punctuation Removal*

Pada tahap *punctuation removal* dilakukan penghapusan tanda baca,angka,dan simbol.

4. *Stemming*

Pada tahap *stemming* mengubah kata-kata menjadi kata dasar dengan cara menghapus imbuhan.

5. *Stopword Removal*

Pada tahap *stopword removal* dilakukan penghilangan kata-kata umum yang tidak relevan, seperti kata "dari", "adalah", "atau", "sebuah", dan lain-lain dalam dokumen bahasa Indonesia.

2.2.3. Penyusunan Kembali kalimat (*Recreate Sentences*)

Recreate sentences adalah tahap menggabungkan seluruh kata pada kalimat asal yang telah mengalami *preprocessing*, sehingga membentuk sebuah kalimat kembali. Contohnya dapat dilihat padad Tabel 2 berikut ini, yang diambil dari artikel pada dataset, dengan judul "*Ini Dia Manfaat Mengonsumsi Jeruk*" .

Tabel 2. Contoh Proses *Recreate Sentence* (Penggabungan Kalimat)

Kalimat awal	Hasil *preprocessing*
1. Jakarta-Indonesia kaya akan beragam jenis jeruk lokal, contohnya jeruk Sambas yang berasal dari Pontianak.	1. jakarta indonesia kaya agam jenis jeruk lokal contoh jeruk sambas asal pontianak
2. Buah yang berwarna orange ini kaya dengan vitamin c dan serat yang dapat memberikan manfaat bagi tubuh.	2. buah warna orange kaya vitamin c serat manfaat tubuh
3. Menurut Kepala Dinas Pertanian, Musanif yang ditemui dalam acara puncak hari Kesehatan Nasional (HKN) yang ke-53 di Jakarta, Minggu (12/11/2017), jeruk seperti Sambas dapat membantu pemenuhan gizi seseorang.	3. kepala dinas tani musanif temu acara puncak sehat nasional hkn ke53 jakarta minggu 12112017 jeruk sambas bantu penuh gizi
4. Konsumsi buah seperti jeruk, menurutnya bisa membantu memenuhi kebutuhan gizi masyarakat.	4. konsumsi buah jeruk turut bantu penuh butuh gizi masyarakat
5. Secara umum jeruk memang mampu memberikan manfaat bagi kesehatan tubuh, seperti penjelasan berikut ini: Untuk menurunkan risiko stroke Menurut American Heart Association, makan dalam jumlah yang lebih tinggi dari senyawa yang ditemukan dalam buah jeruk dapat menurunkan risiko stroke iskemik bagi perempuan.	5. jeruk manfaat sehat tubuh jelas turun risiko stroke american heart association makan senyawa temu buah jeruk turun risiko stroke iskemik perempuan
6. Stroke iskemik terjadi pada sel-sel otak yang mengalami kekurangan oksigen dan nutrisi yang disebabkan penyempitan atau penyumbatan pada pembuluh darah (arteriosklerosis).	6. stroke iskemik selsel otak alami kurang oksigen nutrisi sebab sempit sumbat buluh darah arteriosclerosis
7. Mengkonsumsi jeruk dalam jumlah yang tinggi memiliki risiko 19% lebih rendah terkena stroke iskemik ketimbang wanita yang mengkonsumsi sedikit.	7. konsumsi jeruk milik risiko 19 rendah kena stroke iskemik ketimbang wanita konsumsi
8. Mencegah Kanker Menurut sebuah studi yang dipublikasikan dalam American Journal of Epidemiology, mengkonsumsi manfaat pisang, jeruk dan jus jeruk dalam dua tahun pertama kehidupan dapat mengurangi risiko pengembangan leukimia.	8. cegah kanker studi publikasi american journal of epidemiology konsumsi manfaat pisang jeruk jus jeruk hidup kurang risiko kembang leukimia
9. Sebagai sumber yang sangat baik dari manfaat antioksidan, vitamin C dalam jeruk juga dapat membantu memerangi pembentukan radikal bebas yang diketahui sebagai penyebab kanker.	9. sumber manfaat antioksidan vitamin c jeruk bantu rang bentuk radikal bebas sebab kanker
10. Jumlah vitamin C yang diperlukan yang dikonsumsi untuk tujuan terapeutik kanker dan serat tinggi dari buah-buahan dan sayur-sayuran berkaitan dengan penurunan resiko kanker kolorektal.	10. vitamin c konsumsi tuju terapeutik kanker serat buahbuahan sayursayuran kait turun resiko kanker kolorektal
11. Buah Jeruk membantu menjaga sistem imun Jeruk merupakan sumber vitamin C yang paling tinggi dibandingkan dengan jenis buah-buahan lainnya.	11. buah jeruk bantu jaga sistem imun jeruk sumber vitamin c banding jenis buahbuahan
12. Satu buah jeruk berukuran sedang dapat memenuhi sekitar 72% dari kebutuhan harian vitamin C.Vitamin C merupakan antioksidan yang dapat melindungi tubuh dari kerusakan yang diakibatkan oleh radikal bebas.	12. buah jeruk ukur penuh 72 butuh hari vitamin c vitamin c antioksidan lindung tubuh rusa akibat radikal bebas
13. Vitamin C pada jeruk juga efektif untuk meningkatkan sistem kekebalan tubuh.	13. vitamin c jeruk efektif tingkat sistem kebal tubuh

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

Tahap berikutnya menentukan kemiripan antar kalimat di dalam dokumen dalam bentuk matriks (*similarity matrix*). *Similarity* dihitung melalui pengukuran jarak antar vektor masing-masing kalimat, dengan menggunakan pembobotan TF.IDF. Pada dasarnya, bobot TF-IDF merupakan ukuran statistik untuk mengukur tingkat kepentingan kata dalam sebuah dokumen [11]. Tingkat kepentingan meningkat jika sebuah kata muncul beberapa kali dalam sebuah dokumen tetapi sama dengan frekuensi kemunculan kata tersebut dalam kumpulan dokumen**.** *Term frequency* (TF*)* termasuk pengukuran paling sederhana dalam metode pembobotan. Pada metode ini, tiap-tiap *term* mempunyai tingkat kepentingan sesuai jumlah kemunculan dalam teks dokumen (setelah semua *stopword* dihilangkan), semakin sering muncul, maka kata tersebut semakin penting [12].

doi: https://doi.org/10.37859/coscitech.v3i3.4300

$$IDF_t = log\ (\frac{N}{DF_t}) \quad (1)$$

Sedangkan IDF adalah *Inverse document frequency*, yang secara definisi adalah menyatakan seberapa penting suatu *term* atau token berdasarkan banyaknya dokumen yang mengandung kata tersebut di dalam korpus. Dalam kasus peringkas dokumen ini, koleksi (korpus) adalah artikel, dengan kalimat sebagai dokumen di dalam korpusnya, karena kita mengevaluasi setiap dokumen secara mandiri (terpisah). *IDF* dihitung dengan persamaan (1), dengan *N* merupakan total seluruh kalimat dalam 1 korpus (dalam kasus peringkas dokumen, korpus adalah artikel dokumen yang diringkas), sedangkan DF_t ialah *document frequency* untuk token ke-*t* di dalam kalimat, menyatakan jumlah kalimat yang mengandung *token* ke-*t* . Perkalian bobot *TF* dan bobot *IDF* menjadi bobot untuk setiap token ke-*t* dihitung menggunakan persamaan (2) berikut:

$$W_t = TF_t * IDF_t \quad (2)$$

dengan *IDF* dihitung dari persamaan (1), sedangkan *t* adalah *token* kata (*term*) ke-*t* dari kalimat, untuk *TF* adalah banyaknya *term* pada sebuah kalimat. Perhitungan panjang vektor kalimat di dalam korpus (artikel) menggunakan persamaan (3) berikut, dengan *n* adalah jumlah token kata (*term*) pada kalimat tersebut.

$$|S| = \sqrt{(W_1{}^2) + (W_2{}^2) + \cdots + (W_n{}^2)} = \sqrt{\sum_{t=1}^{n} W_t} \quad (3)$$

2.4. Pembentukan Matriks Kemiripan (*Similarity Matrix*)

Pada algoritma *lexrank* formula *similarity* menggunakan *idf-modified-cosince*. *Idf-modified-cosine* digunakan untuk menghitung bobot dari *term* pada tiap pasang kalimat menggunakan pembobotan *TF.IDF*. Metode *cosine similarity* digunakan untuk menghitung *similarity* antar dua vektor atau lebih [13]. Metode *cosine similarity* diterapkan untuk menghitung tingkat kesamaan antar dokumen tetapi tidak menyertakan frekuensi kata (*term*) [14]. Dengan pembobotan *TF.IDF* ini dapat memberikan nilai bobot pada setiap kata dalam dokumen yang dapat membantu *cosine similarity* untuk memproses kata secara maksimal. Kelebihan dari algoritma *cosine similarity* adalah tidak terpengaruh pada panjang pendeknya suatu dokumen dan memiliki tingkat akurasi yang tinggi. Berikut rumus *cosine similarity* antara vektor kalimat *A* dan kalimat *B*, dihitung dengan persamaan (4) berikut,

$$Similarity = cos(\theta) = \frac{A.B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\ \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (4)$$

dengan *A* dan *B* adalah nilai vektor *A* dan vektor *B*, sedangkan *A.B* merupakan *dot product* antara setiap komponen (bobot kata) vektor A dan vektor B. untuk $|A|$ merupakan panjang vektor *A*, dan untuk $|B|$ adalah panjang vektor *B*, yang masing-masing dihitung dengan persamaan (3).

2.5. Representasi *Graph*

Pada tahap pembentukan graf, simpul (*vertex*) menampilkan kalimat-kalimat yang ada dalam dokumen, sedangkan sisi (*edge)* menampilkan *similarity* antara dua kalimat yang terhubung. Graf terdiri atas titik simpul (*vertex*) yang tidak berurutan, dihubungkan oleh titik- titik sisi (*edge*). Pemodelan Graf menggambarkan dokumen (artikel) yang dipecah ke dalam unit-unit kalimat sebagai simpul dan menambahkan garis penghubung berdasarkan *similarity* antar unit kalimat tersebut. Tingkat pentingnya setiap simpul (kalimat) berdasarkan komposisi dari graf keseluruhan [2].

2.6. Perangkingan Kalimat (*Sentence Ranking*)

Tahap selanjutnya perangkingan kalimat untuk mendapatkan skor akhir dari suatu *vertex*. Semakin besar skor akhir dari suatu simpul, maka semakin banyak informasi di dalamnya. Kalimat yang sudah memperoleh skor, kemudian diberi peringkat berdasarkan skor tertinggi ke skor yang terendah. Persamaan (5) berikut merupakan rumus tahap perangkingan kalimat menggunakan algoritma *lexrank.*

$$p(u) = \frac{d}{N} + (1-d) \sum_{v \in adj[u]} \frac{p(v)}{deg(v)} \quad (5)$$

dengan *N* adalah jumlah *node* atau *vertex* yang ada dalam graf, sedangkan *d* adalah *damping factor* (nilai default=0,85), dan *u* adalah matriks persegi yang semua elemennya =1/*n*. Kalimat yang terpilih sebagai ringkasan diambil dari sejumlah kalimat dengan *rank* teratas sesuai dengan tingkat kompresi (*compression rate*) ringkasan yang diinginkan.

doi: https://doi.org/10.37859/coscitech.v3i3.4300

2.7. Evaluasi

Untuk pengukuran performa sistem, digunakan *ROUGE* (*Recall-Oriented Underresearch for Gisting Evaluation) scoring*, yang diperoleh dengan membandingkan hasil ringkasan teks oleh sistem (metode) dengan hasil ringkasan teks yang dibuat oleh manusia (*human annotator*). Penelitian ini, akan menggunakan *ROUGE-N*, dan *ROUGE-L* yang terdiri atas *precision, recall, dan F-measure* sebagai *scoring* untuk mengukur kinerja sistem, mengikuti banyak penelitian di bidang peringkasan dokumen seperti pada [15].

2.7.1. ROUGE-N

ROUGE-N membandingkan n-grams jumlah kata yang sesuai antara ringkasan sistem dan ringkasan manual. Jumlah n -gram yang digunakan beragam, umumnya n={1,..,4}. Tetapi, yang paling banyak digunakan di dalam penelitian adalah n-gram dengan jumlah n=1 (*ROUGE-1*) dan n=2 (*ROUGE-2*). Persamaan (6-7) merupakan rumus *recall* dan *precision* untuk *ROUGE-1* dan *ROUGE-2*.

$$ROUGE-1\ recall = \frac{jumlah\ unigram\ kata\ yang\ sama}{Total\ kata\ di ringkasan\ manual} \quad (6)$$

$$ROUGE-1\ precision = \frac{jumlah\ unigram\ kata\ yang\ sama}{Total\ kata\ di ringkasan\ sistem} \quad (7)$$

$$ROUGE-2\ recall = \frac{jumlah\ bigram\ kata\ yang\ sama}{Total\ kata\ di ringkasan\ manual} \quad (8)$$

$$ROUGE-2\ precision = \frac{jumlah\ bigram\ kata\ yang\ sama}{Total\ kata\ di ringkasan\ sistem} \quad (9)$$

2.7.2. ROUGE-L

ROUGE-L bekerja dengan mencocokan seluruh susunan kata terpanjang yang sama yang dikenal dengan *longest common subsequence* (LCS) antara ringkasan sistem dan ringkasan manual. Berikut merupakan rumus *precision,recall* untuk *ROUGE-L*.

$$ROUGE-L\ recall = \frac{LCS(sistem\ ,\ manual)}{Total\ kata\ di ringkasan\ manual} \quad (10)$$

$$ROUGE-L\ precision = \frac{LCS(si\quad tem,\ manual)}{Total\ kata\ di ringkasan\ sistem} \quad (11)$$

2.7.3. F-Measure

F-Measure adalah hasil kombinasi antara nilai *recall* dan *precission* untuk menghitung akurasi suatu sistem. Komposisi berimbang antara *precision* dan *recall* disebut dengan *F1-measure*, yang dihitung dengan rumus (12) berikut ini.

$$F1-Measure\ = 2 * \frac{precision * recall}{precision + recall} \quad (12)$$

3. HASIL DAN PEMBAHASAN

3.1. *Set up* Eksperimen

Dataset pada penelitian ini berupa korpus sebanyak 300 artikel, yang dibagi atas 2 bagian, dengan komposisi 50:50. Sebanyak 150 artikel digunakan untuk pengembangan sistem dan mencari *model summarizer* terbaik dari *LexRank* dalam penelitian ini. Sisanya 150 artikel lainnya digunakan untuk pengujian (sebagai *data testing*). Tingkat kompresi (*compression rate*) yang dipilih adalah 30% dan 50% dari jumlah kalimat setiap artikel (dokumen). Model dipilih dengan cara mencari bagian *score LexRank* mana yang paling baik untuk menghasilkan ringkasan yang paling mendekati ringkasan manual (*gold standard*), di antara kelompok kalimat dengan skor tertinggi, kelompok kalimat dengan skor terendah, atau kelompok pertengahan.

Model yang terpilih, kemudian digunakan untuk menguji data testing, dan diukur performanya menggunakan *ROUGE score*. Tabel 3 berikut merupakan perbandingan jumlah kata rata-rata setiap artikel pada dataset (*average word each document*) dan jumlah kata rata-rata ringkasan oleh sistem (*average word each summary*) yang dinamakan *average compression ratio* (kolom terakhir). Statistik ini mengikuti data yang dihitung dan dianalisa oleh penelitian tentang ringkasan yang sejenis seperti pada[16].

doi: https://doi.org/10.37859/coscitech.v3i3.4300

Tabel 3. Hasil Perhitungan Rata-rata *Compression Rate* (*average compression ratio*)

No	*Compression Rate*	Rata-rata jumlah kata per artikel	Rata-rata jumlah kata per ringkasan gold standard	Rata-rata rasio *Compression Rate*
1	30%	399	131	**67,1%**
2	50%	399	215	**46,1%**

3.2. *LexRank Similarity*

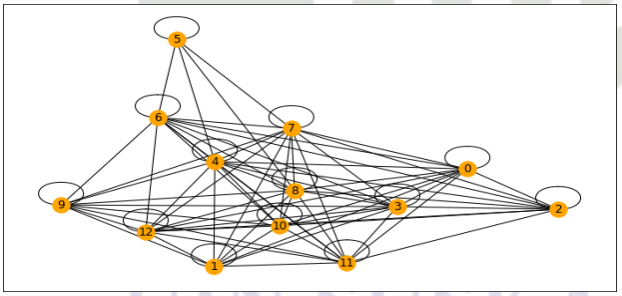
Setelah kalimat-kalimat di dalam dokumen diproses sebagaimana pada penjelasan bagian 2.2 dan 2.3, hasil matirks *similarity* antar kalimat dihitung sebagaimana diterangkan pada bagian 2.4, disajikan pada Tabel 4 berikut ini. Skala logaritmik digunakan untuk mencegah hasil berupa angka yang sangat kecil dari perkalian vektor dengan *sparsity* yang sangat besar dari *Bag of Word*, karena vektor kalimat memiliki sangat banyak nilai nol untuk kata-kata yang tidak terkandung dalam kalimat, dengan S adalah simbol dari vektor kalimat, dan T adalah *vector transpose* dari kalimat 1 sampai 13 di dalam dokumen yang dijadikan contoh.

Tabel 4. Similarity antar kalimat menggunakan *cosine similarity* dari norm vektor dalam skala logaritmik

T	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
T1	0,227	0,675	0,041	0,013	0,015	0	0,01	0,016	0.011	0	0,046	0,009	0,015
T2	0,029	0,277	0	0,02	0.038	0	0	0.015	0.035	0.039	0.031	0.054	0.045
T3	0,041	0	0,277	0,58	0,03	0	0,004	0,006	0,016	0	0,021	0,015	0,006
T4	0,013	0,02	0,058	0,277	0,02	0	0,042	0,025	0,027	0,017	0,051	0,064	0,01
T5	0,015	0,038	0,03	0,02	0,277	0,032	0,067	0,051	0,019	0,028	0,025	0,023	0,025
T6	0	0	0	0	0,032	0,277	0,029	0,015	0,021	0	0	0	0
T7	0,01	0	0,004	0,042	0,067	0,029	0,277	0,043	0,006	0,025	0,011	0,005	0,007
T8	0,016	0,015	0,006	0,025	0,051	0,015	0,043	0,277	0,036	0,035	0,018	0,007	0,012
T9	0,011	0,035	0,016	0,027	0,019	0,021	0,006	0,036	0,277	0,044	0,068	0,088	0,023
T10	0	0,039	0	0,017	0,028	0	0,025	0,035	0,044	0,277	0,031	0,016	0,012
T11	0,046	0,031	0,021	0,051	0,025	0	0,011	0,018	0,068	0,031	0,277	0,04	0,065
T12	0,009	0,054	0,015	0,064	0,023	0	0,005	0,007	0,088	0,016	0,04	0,277	0,048
T13	0,015	0,045	0,006	0,01	0,025	0	0,007	0,012	0,023	0,012	0,065	0,048	0,277

3.3. *LexRank Graph*

Untuk menggambarkan kedekatan antar kalimat, graf dari kalimat-kalimat (*vertex*) penyusun dokumen ditampilkan dalam representasi graf sebagaimana gambar 2 berikut ini. Dari gambar contoh kasus di bawah ini, terlihat bahwa kalimat ke-8 memiliki kedekatan dengan kalimat 10, lalu dengan kalimat 7 dan 4. Dari gambar tersebut, dapat pula disebutkan bahwa kalimat 8 menjadi pusat topik dari artikel tersebut.



Gambar 2. Representasi Graph

3.4. Pembentukan Ringkasan (*Summary Formation*)

Tahap pembentukan ringkasan yaitu proses penggabungan dari kalimat-kalimat yang telah dirangking sebelumnya. Pada penelitian ini diguaka *compression rate* 30% dan 50% dari teks asli yang ada pada setiap dokumen/artikel. Proses pembentukan ringkasan dengan cara mengelompokkan kalimat-kalimat menjadi tiga bagian yaitu *summary* berdasarkan ranking awal (*top-n*/ n-teratas), *summary* tengah (*med-n*, yang berada di pertengahan ranking), dan *summary* akhir (*bottom-n*/n-terbawah). Berikut salah satu hasil *compression ratio* menggunakan algoritma *lexrank* dengan *compression ratio* 50% sehingga menghasilkan 6 kalimat tiap bagiannya (pembulatan kebawah). Tabel 4 berikut ini adalah hasil pengelompokan ranking yang dimaksud.

doi: https://doi.org/10.37859/coscitech.v3i3.4300

Peringkat	Skor Kalimat	Nomor Kalimat
1	1.4338762179822868	10
2	1.2524404381006213	12
3	1.1978144088621825	0
4	1.1720112229718886	4
5	1.1468066352016875	8
6	1.14343555749768	3
7	1.1407173557032375	11
8	1.094768106387335	7
9	0.9709008059013491	6
10	0.8330317745440715	2
11	0.6695645821752677	1
12	0.5597877246597954	9
13	0.38484517001260077	5

awal
tengah
akhir

Tabel 5. Pengelompokan ranking kalimat untuk kandidat ringkasan

Dari tabel diatas didapat tiga macam ringkasan:

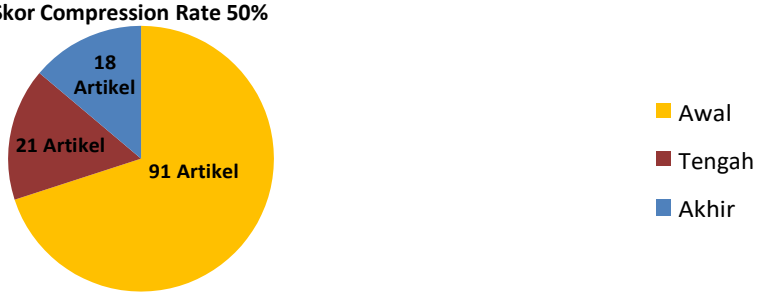
- *summary* awal terdiri dari kalimat 0,3,4,8,10,12,
- *summary* tengah terdiri dari kalimat 3,4,6,7,8,11, dan
- *summary* akhir terdiri dari kalimat 1,2,5,6,7,9.

3.5 Hasil Pengujian Data Model *Compression Rate* 50%

Dari 150 dataset *development* (yang dipakai untuk menentukan *summary model* terbaik) dihitung scoring menggunakan *ROUGE score* untuk ketiga kandidat *summary*-nya. Nilai *ROUGE-L* terbaik dari masing-masing dokumen ditandai, dan dihitung statistiknya. Metrik yang dipakai untuk pengukuran kinerja terbaik adalah *F-measure*. Tabel 5 berikut merupakan cuplikan hasil dari pencarian posisi skor *lexrank* terbanyak model dengan kompresi 50%, dari artikel No. 1, diperoleh data bahwa *ROUGE-L score* terbaik bila menggunakan kelompok *ranking* akhir. Artikel No.2 terbaik menggunkan kelompok ranking awal, dan artikel No.3 menggunakan kelompok ranking tengah. Secara keseluruhan, distribusinya dapat dilihat pada Gambar 3 di bawah ini.

Tabel 6. Hasil Pengujian Data Model *Compression Rate* 50%

No	Judul Artikel	Skor *LexRank* dipilih	Rouge-1 (%)			Rouge-2(%)			Rouge-L(%)			Max
			R	P	F	R	P	F	R	P	F	
1	7 Situs Berita Online Terpopuler Di Indonesia.	Awal	47,56	62,40	53,98	42,68	53,97	47,66	47,56	62,40	53,98	
		Tengah	54,88	69,23	61,22	44,35	59,89	50,96	54,88	69,23	61,22	Akhir
		Akhir	75,00	67,96	71,30	55,23	56,65	55,93	74,39	67,40	**70,72**	
2	Aksi Menarik Cristiano Ronaldo Berolahraga Sembari Mengasuh Anak	Awal	81,93	80,95	81,44	72,73	74,23	73,47	81,93	80,95	**81,44**	
		Tengah	24,10	31,75	27,40	12,12	16,44	13,95	24,10	31,75	27,40	Awal
		Akhir	40,96	52,31	45,95	26,26	36,11	30,41	40,96	52,31	45,95	
3	Algoritma di balik aplikasi perencanaan trip wisata .	Awal	50,72	46,05	48,28	39,56	40,45	40,00	49.28	44,74	46,90	
		Tengah	66,67	56,10	60,93	59,34	53,47	56,35	66,67	56,10	**60,93**	Tengah
		Akhir	68,12	47,47	55,95	59,34	47,37	52,68	68,12	47,47	55,95	
148	Satgas Covid-19 Minta Masyarakat Adaptif dengan Perubahan Aturan Perjalanan.	Awal	60,66	82,22	69,81	60,96	77,39	68,20	60,66	82,22	**69,81**	
		Tengah	43,44	55,21	48,62	34,93	43,97	38,93	42,62	54,17	41,71	Awal
		Akhir	46,72	55,88	50,89	38,36	47,46	42,42	46,72	55,88	50,89	
149	Satuan Reserse Kriminal Polres Dumai Berhasil Menangkap 4 Pelaku Ilegal Logging.	Awal	46,15	59,02	51,80	40,40	51,95	45,45	46,15	59,02	51,80	
		Tengah	48,08	60,48	53,57	39,39	53,79	45,48	46,15	58,06	51,43	Akhir
		Akhir	65.38	90,27	75,84	58,59	87,22	70,09	65,38	90,27	**75,84**	
150	Seorang Wn Inggris Tersandung Kasus Narkoba Di Pekanbaru.	Awal	64,52	77,52	70,42	55,71	70,93	62,40	64,52	77,52	70,42	
		Tengah	65,81	77,27	71,08	55,25	70,76	62,05	65,81	77,27	**71,08**	Tengah
		Akhir	58,71	66,91	62,54	44,75	56,32	49,87	58,71	66,91	62,54	



Gambar 3. Skor *LexRank* Maksimal *Compression Rate* 50%

doi: https://doi.org/10.37859/coscitech.v3i3.4300

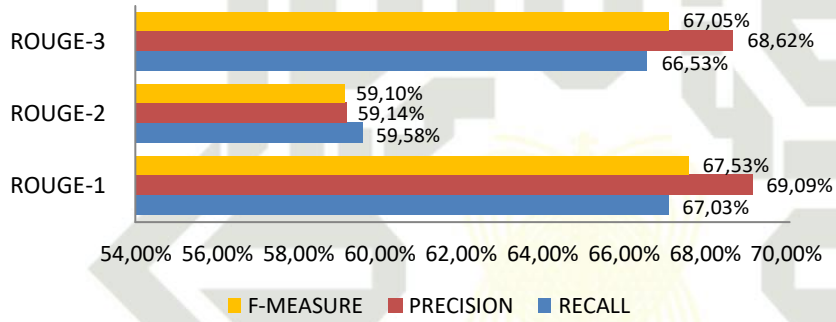
Berdasarkan skor pada Gambar 3 diatas, diperoleh posisi *ROUGE-L* tertinggi berada pada posisi ringkasan awal sebanyak 91 artikel. Model ini dipakai untuk system summary berbasis *LexRank* yang digunakan dalam penelitian ini untuk membangkitakn secara otomatis ringkasa terhadap data test (150 artikel lainnya).

3.6 Hasil Pengujian Data Testing Kompresi 50%

Tabel 6. berikut merupakan hasil pengujian terhadap data testing, dengan kompresi 50% menggunakan 150 artikel menggunakan skor *lexrank* kelompok ranking awal.

Tabel 7. Hasil Pengujian Data Testing Kompresi 50%

No	Judul Artikel	Rouge-1 (%)			Rouge-2(%)			Rouge-L(%)		
		R	P	F	R	P	F	R	P	F
1	(HUT) Harian Pekanbaru MX	78,00	75,97	76,97	71,14	68,24	69,66	78,00	75,97	76,97
2	5 Kondisi Kesehatan Gigi dan Mulut yang Menandakan Pikiran Sedang Stres	53,33	70,33	60,66	47,44	65,49	55,02	52,50	69,23	59,72
3	5 Penemuan Muslim Yang Mengubah Dunia	64,94	62,86	63,88	54,94	53,27	54,09	63,84	61,79	62,79
...										
148	Yahoo Hentikan Layanan di Tiongkok	67,31	80,46	73,30	55,79	64,66	60,00	67,31	80,46	73,30
149	Yuk Kenali Produk Vegan	54,24	57,14	55,65	51,39	50,68	51,03	52,54	55,36	53,91
150	Yuk Mengenal Masjid Quba	92,19	75,16	82,81	87,65	68,66	77,00	92,19	75,16	82.81
	RATA-RATA	**67,03**	**69,09**	**67,53**	**59,74**	**59,58**	**59,14**	**66,53**	**68,62**	**67,05**



Gambar 4. Visualisasi Hasil Testing Kompresi 50%

Secara statistik, hasil *ROUGE Score* yang diperoleh untuk ringkasan menggunakan *score LexRank* kelompok awal terhadap data testing adalah seperti terlihat pada Gambar 4. Terlihat hasil *ROUGE*-1 dan *ROUGE-L* yang amat baik melebihi angka 50% untuk *compression rate* 50%.

3.7 Hasil Pengujian Data Model Kompresi 30%

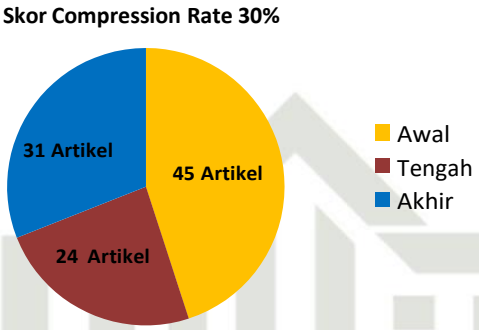
Tabel 7. berikut merupakan hasil dari pencarian skor *lexrank* dari data model dengan kompresi 30% menggunakan 150 artikel. Kita perlu membuat ulang model, karena rate kompresi yang berubah, memungkinkan *scoring ROUGE* juga akan berubah. Dari hasil penelusuran terhadap data model (*development*), diperoleh hasil bahwa skor *ROUGE* terbaik berada pada kelompok ranking awal, namun dengan jumlah yang sudah hamper sama dengan kelompok ranking akhir (Gambar 5). Bukan tidak mungkin, untuk bentuk dokumen lain atau dokumen-dokumen yang lebih panjang dari data yang dipakai pada penelitian ini, hasil ini akan berbeda.

Tabel 8. Hasil Pengujian Data Model Kompresi 50%

No	Judul Artikel	Skor *LexRank* dipilih	Rouge-1 (%)			Rouge-2(%)			Rouge-L(%)			Max
			R	P	F	R	P	F	R	P	F	
1	7 Situs Berita Online Terpopuler Di Indonesia.	Awal	40,34	56,47	47,06	35,26	46,61	40,15	40,34	56,47	**47,06**	
		Tengah	40,34	53,93	46,15	28,21	38,26	32,47	40,34	53,93	46,15	Awal
		Akhir	43,70	50,49	46,85	23,08	27,07	24,91	42,02	48,54	45,05	
2	Aksi Memarik Cristiano Ronaldo Berolahraga Sembari Mengasuh Anak	Awal	68,00	68,00	68,00	50,00	53,70	51,79	66,00	66,00	**66,00**	
		Tengah	22,00	25,00	23,40	5,17	6,12	5,61	22,00	25,00	23,40	Awal
		Akhir	66,00	47,83	55,46	51,72	37,97	43,80	66,00	47,83	55,46	
3	Algoritma di balik aplikasi perencanaan trip wisata	Awal	85,19	41,07	55,42	76,67	38,33	51,11	85,19	41,07	55,42	
		Tengah	66,67	54,55	60,00	53,33	44,44	48,48	66,67	54,55	**60,00**	Tengah
		Akhir	33,33	18,37	23,68	20,00	10,53	13,97	29,63	16,33	21,05	
...												
148	Satgas Covid-19 Minta	Awal	41,89	55,36	47,69	34,15	43,08	38,10	41,89	55,36	**47,69**	

doi: https://doi.org/10.37859/coscitech.v3i3.4300

	Masyarakat Adaptif dengan Perubahan Aturan Perjalanan.	Tengah	9,46	15,56	11,76	1,00	1,00	1,00	9,46	15,56	11,76	Awal
		Akhir	36,49	57,45	44,63	30,49	48,08	37,31	36,49	57,45	44,63	
149	Satuan Reserse Kriminal Polres Dumai Berhasil Menangkap 4 Pelaku Ilegal Logging.	Awal	23,40	27,85	25,43	15,79	19,15	17,31	22,34	26,58	24,28	
		Tengah	40,43	43,18	41,76	31,58	36,36	33,80	40,43	43,18	41,76	Akhir
		Akhir	46,81	53,01	49,72	40,35	45,10	42,59	45,74	51,81	**48,59**	
150	Seorang Wn Inggris Tersandung Kasus Narkoba Di Pekanbaru.	Awal	53,33	63,64	58,03	39,55	47,75	43,21	51,43	61,36	55,96	
		Tengah	54,29	74,03	62,64	43,28	63,04	51,33	54,29	74,03	**62,64**	Tengah
		Akhir	36,19	48,72	41,53	21,64	30,21	25,22	36,19	48,72	41,53	



Gambar 5. Skor *LexRank* Maksimal *Compression Rate* 30%

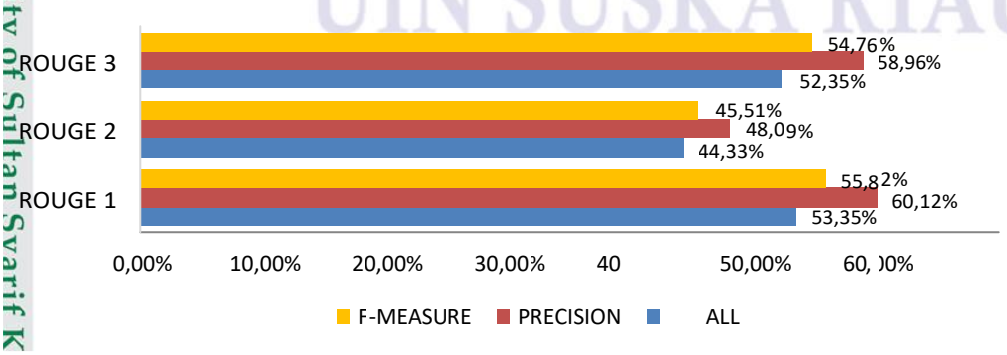
Berdasarkan diagram pada gambar 5 diatas, maka model *summary* yang dipakai adalah ranking bagian awal. Model ini masih sama dengan model *summary* untuk compression rate 50%

3.3 Hasil Pengujian Data Testing Kompresi 30%

Tabel 9. memperlihatkan cuplikan hasil pengujian terhadap data testing yang terdiri atas 150 dokumen. Secara statistik,hasil yang dicapai sebagaimana pada Gambar 6, menunjukkan hasil *ROUGE* score yang masih baik, yaitu di atas 50% untuk *compression rate* 30%. Hasil ini bila dibandingkan dengan beberapa penelitian di bidang *text summarization* yang berbasis ekstraksi kalimat, termasuk ke dalam hasil yang baik, karena rata-rata peringkas dokumen memiliki *f-measure* dari *ROUGE-L* tidak sampai 50% [6], [7]. Namun demikian, *scoring* pada penelitian *text summarization* perlu dibandingkan pada penggunaan dataset yang sama, agar perbandingannya dapat lebih terukur dengan valid.

Tabel 3. Hasil Pengujian Data Testing Kompresi 30%

No	Judul Artikel	Rouge-1 (%)			Rouge-2(%)			Rouge-L(%)		
		R	P	F	R	P	F	R	P	F
1	(HUT) Harian Pekanbaru MX	36,49	57,45	44,63	30,49	48,08	37,31	36,49	57,45	44,63
2	5 Kondisi Kesehatan Gigi dan Mulut yang Menandakan Pikiran Sedang Stres	65,53	72,97	69,05	59,09	63,77	61,34	65,53	72,97	69,05
3	5 Penemuan Muslim Yang Mengubah Dunia	41,10	48,39	44,44	35,23	40,79	37,80	39,73	46,77	42,96
...										
148	Yahoo Hentikan Layanan di Tiongkok	22,22	29,63	25,40	2,38	3,03	2,67	13,89	18,52	15,87
149	Yuk Kenali Produk Vegan	34,78	44,44	39,02	22,22	27,27	24,49	30,43	38,89	34,15
150	Yuk Mengenal Masjid Quba	42,68	33,65	37,63	28,85	22,22	25,10	40,24	31,73	35,48
	RATA-RATA	**53,35**	**60,12**	**55,82**	**44,33**	**48,09**	**45,51**	**52,35**	**58,96**	**54,76**



Gambar 6. Visualisasi Hasil Testing Kompresi 30%

doi: https://doi.org/10.37859/coscitech.v3i3.4300

4. KESIMPULAN

Berdasarkan penelitian yang diakukan, disimpulkan beberapa hal-hal sebagai berikut:

1. Algoritma *LexRank* berhasil diterapkan dalam peringkasan teks otomatis dokumen bebahasa Indonesia yang secara umum memiliki skor *ROUGE* yang kompetitif dalam tugas peringkasan dokumen.
2. Untuk dataset yang telah dikembangkan dalam penelitian ini, *lexrank* dengan modifikasi seleksi ranking, pada pengujian dengan *Compression Rate* 50% menghasilkan skor *f-measure* 67,53% pada *ROUGE-1*, 59,10% pada *ROUGE-2*, dan 67,05% pada *ROUGE-L*. Sedangkan untuk pengujian *Compression Rate* 30% diperoleh rata-rata *f-measure* 55,82% pada *ROUGE-1*, 45,51% pada *ROUGE-2*, dan 54,76% pada *ROUGE-L*. Hasil ini lebih baik jika dibandingkan dengan beberapa penelitian metode ekstraktif lainnya.
3. Tingkat kompresi yang digunakan sangat berpengaruh terhadap hasil pengujian performa pada metode yang digunakan. Semakin kecil tingkat kompresi yang digunakan, maka hasil yang dicapai pada umumnya akan semakin kecil, karena peluang pemilihan kalimat oleh metode tidak cocok dengan pemilihan kalimat oleh manusia menjadi semakin besar.
4. Saran untuk penelitian selanjutnya, dapat menggunakan dataset ini untuk metode ringkasan yang lain, agar dapat dibandingkan hasilnya. Penggunaan dataset yang standard dalam tugas peringkasan dokumen di dunia (*shared task*) juga dapat dilakukan untuk menguji metode ini di antara hasil-hasil yang telah pernah dilaporkan.

DAFTAR PUSTAKA

[1] G. Mandar and G. Gunawan, "Peringkasan dokumen berita bahasa indonesia menggunakan metode cross latent semantic analysis," *Regist. J. Ilm. Teknol. Sist. Inf.*, vol. 3, no. 2, pp. 94–104, 2017, doi: 10.26594/register.v3i2.1161.

[2] A. Fauzi, "Penerapan Algoritma Text Mining dan Lexrank dalam Meringkas Teks Secara Otomatis," vol. 1, no. 2, pp. 65–72, 2022, [Online]. Available: https://ejurnal.seminar-id.com/index.php/bulletinds%0APenerapan.

[3] K. U. Syaliman and Yuliska, " Literatur Review Terhadap Metode , Aplikasi dan Dataset Peringkasan Dokumen Teks Otomatis untuk Teks Berbahasa Indonesia," *IT J. Res. Dev.*, vol. 5, no. 1, pp. 19–31, 2020.

[4] D. R. Radev and G. Erkan, "LexRank : Graph-based Centrality as Salience in Text Summarization," *J. Artif. Intell. Res.*, vol. 22, no. 1, pp. 457–479, 2004, [Online]. Available: https://arxiv.org/abs/1109.2128.

[5] A. Munshi, A. Mehra, and A. Choudhury, "LexRank Algorithm: Application in Emails and Comparative Analysis," *Int. J. New Technol. Res.*, vol. 7, no. 5, pp. 34–38, 2021, doi: 10.31871/ijntr.7.5.9.

[6] Y. M. Sari and N. S. Fatonah, "Peringkasan Teks Otomatis pada Modul Pembelajaran Berbahasa Indonesia Menggunakan Cross Latent Semantic Analysis (CLSA)," *JEPIN (Jurnal Edukasi dan Penelit. Inform.*, vol. 7, no. 2, pp. 153–159, 2021.

[7] S. Riyadi and I. Samiati, "Pengenalan algortima dalam model peringkasan teks untuk mempercepat pekerjaan akademik," *JGMM*, vol. 1, no. 2, pp. 19–26, 2021.

[8] A. Samuel and D. K. Sharma, "Modified LexRank for Tweet Summarization," *Int. J. Rough Sets Data Anal.*, vol. 3, no. 4, pp. 79–90, 2016, doi: 10.4018/ijrsda.2016100106.

[9] J. Pragantha, C. V. M, and Eris, " Penerapan Algoritma Textrank Untuk Automatic Summarization Pada Dokumen Berbahasa I ndonesia," *J. Ilmu Tek. dan Komput.*, vol. 1, no. 1, pp. 71–78, 2017.

[10] P. G. Somantri, A. Komarudin, and R. Ilyas, "Peringkasan Teks Otomatis Berita Berdasarkan Klasifikasi Kalimat Menggunakan Sup port Vector Machine," in *Prosiding SNATIF*, 2018, pp. 57–62.

[11] Ash Shiddicky and Surya Agustian, " Analisis Sentimen Masyarakat Terhadap Kebijakan Vaksinasi Covid-19 pada Media Sosial Twitter menggunakan Metode Logistic Regression," *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 3, no. 2, pp. 99–106, 2022, doi: 10.37859/coscitech.v3i2.3836.

[12] R. Gunawan, R. Septiadi, F. Apri Wenando, H. Mukhtar, and Syahril, " K-Nearest Neighbor (KNN) untuk Menganalisis Sentimen terhadap Kebijakan Merdeka Belajar Kampus Merdeka pada Komentar Twitter," *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 3, no. 2, pp. 152–158, 2022, doi: 10.37859/coscitech.v3i2.3841.

[13] M. M. Sya'bani and R. Umilasari, " Penerapan Metode Cosine Similarity dan Pembobotan TF/IDF pada Sistem Klasifikasi Sinopsis Buku di Perpustakaan Kejaksaan Negeri Jember," *JUSTINDO (Jurnal Sist. dan Teknol. Inf. Indones.*, no. Vol 3, No 1 (2018): JUSTINDO, pp. 31 –42, 2018, [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/JUSTINDO/article/view/2345.

[14] A. Apriani, H. Zakiyudin, and K. Marzuki, " Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF System Penerimaan Mahasiswa Baru pada Kampus Swasta," *J. Bumigora Inf. Technol.*, vol. 3, no. 1, pp. 19–27, 2021, doi: 10.30812/bite.v3i1.1110.

[15] M. A. Ubaidillah, I. B. G. Dwidasmara, and A. Muliant ara, "Peringkas Teks Otomatis Berita Online Menggunakan Metode Cross Latent Semantic Analysis & Cosine Similarity," *J. Elektron. Ilmu Komput. Udayana*, vol. 9, no. 1, pp. 105–113, 2020.

[16] Yuliska and U. K. Syaliman, " Peringkasan dokumen teks otomatis berdasarkan sebuah kueri menggunakan bidirectional long short term memory network automatic text document summary based on a query using bidirectional long short term memory network," *J. Inf. Technol. Comput. Sci.*, vol. 5, no. 2, pp. 65–71, 2022.

doi: https://doi.org/10.37859/coscitech.v3i3.4300