

Peringkasan Teks Otomatis Artikel Berbahasa Indonesia Menggunakan Algoritma *Textrank*

Faizah Husniah¹⁾, Surya Agustian²⁾, & Iis Afrianty³⁾

^{1,2,3} Teknik Informatika, Fakultas Sains & Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau,
Jl.H.R. Soebrantas No. 155 KM.1,5 Simpang Baru Panam Pekanbaru, 0761-858832,

Website: www.uin-suska.ac.id,

E-mail:

11850120304@students.uin-suska.ac.id, surya.agustian@uin-suska.ac.id, iis.afrianty@uin-suska.ac.id

Abstrak

Banyaknya informasi yang terdapat pada media online salah satunya adalah artikel. Artikel pada umumnya berisikan kumpulan kalimat yang panjang sehingga pembaca membutuhkan waktu yang lebih banyak untuk memperoleh intisari dari artikel. Sistem peringkasan teks otomatis membantu pengguna mengetahui isi artikel tanpa harus membaca keseluruhan dokumen, sehingga dapat menghemat waktu. Penelitian ini menyelidiki algoritma *textrank* untuk peringkasan dokumen. Algoritma *textrank* akan menghitung skor dari setiap kalimat, kemudian memilih sejumlah kalimat sebagai ringkasan. *ROUGE* scoring digunakan untuk menghitung performa sistem berdasarkan ringkasan manual yang dibuat oleh manusia (*human annotator*) sebagai *gold standard*. Hasil pengujian dengan *compression rate* 50% pada pilihan ranking kalimat yang terpilih sebagai ringkasan, menghasilkan nilai pada *ROUGE-1*, *ROUGE-2* dan *ROUGE-L* secara berturut-turut adalah 68.76%, 60.60%, dan 68.29%. Sedangkan untuk *compression rate* 30%, rata-rata pada *ROUGE-1*, *ROUGE-2* dan *ROUGE-L* berturut-turut adalah 45.00%, 31.62%, dan 43.59%. Hasil ini sangat kompetitif dibandingkan dengan score pada penelitian tentang sistem peringkasan dokumen lainnya.

Keyword: peringkasan teks otomatis; *compression rate*; *rouge*; skor kalimat; *textrank*

Abstract

The amount of information contained in online media, one of them is articles. Articles generally contain a long collection of sentences so that readers need more time to get the point of the article. The automatic text summarization system helps users to find out the contents of the article without having to read the entire document, thus saving time. This research to inspect the *textrank* algorithm for document summarization. The *textrank* algorithm will calculate the score of each sentence, then choose the sentence number as a summary. The *ROUGE* scoring is used to calculate system performance based on a human annotator as the gold standard. The results of the test with a *compression level* of 50% on the choice of sentence ratings that were selected as a summary, resulted in values on *ROUGE-1*, *ROUGE-2* and *ROUGE-L* sequentially 68.76%, 60.60%, and 68.29%. As for the *compression rate* 30%, the average *s* at *ROUGE-1*, *ROUGE-2* and *ROUGE-L* are 45.00%, 31.62%, and 43.59%, sequentially. These results are very competitive compared with the scores on the other about research about document summary systems.

Kata kunci: automatic text summarization; *compression rate*; *rouge*; sentences score; *textrank*

1 PENDAHULUAN

Pada saat ini, informasi telah menjadi kebutuhan penting bagi manusia. Kebutuhan akan informasi dapat diperoleh lewat berbagai media, salah satunya media digital [1]. Media digital merupakan media elektronik untuk menyimpan dan menerima informasi dalam bentuk digital. Internet sebagai media digital merupakan solusi untuk mendapatkan informasi [2]. Informasi pada internet salah satunya adalah melalui artikel. Dengan banyaknya informasi yang terdapat pada artikel, maka pembaca terkadang kesulitan untuk mengetahui inti sari dari artikel tersebut, selain itu pembaca harus meluangkan banyak waktu untuk

membaca artikel tersebut, sehingga dibutuhkan sebuah sistem untuk meringkas dokumen secara otomatis [3].

Peringkasan teks otomatis adalah bidang dari *information extraction*, yang meringkas dokumen teks berbentuk digital, namun tetap mempertahankan isi informasinya. *Information extraction* yaitu mengekstraksi informasi terstruktur secara otomatis dari dokumen tidak terstruktur atau semi terstruktur. Terdapat dua klasifikasi metode *text summarization* yaitu peringkasan dengan cara ekstraktif dan peringkasan dengan cara abstraktif [1, 4].

Peringkasan ekstraktif dilakukan dengan mengekstraksi sebagian kalimat penting dari dokumen awal mulanya tanpa mengganti struktur kalimat. Peringkasan abstraktif dilakukan dengan membuat dan

menyusun kalimat baru, yang pada dasarnya merupakan kombinasi informasi dari dokumen yang diringkas [3]. Pada umumnya, peringkasan teks digolongkan ke dalam metode *extractive summarization* bila berdasarkan proses ekstraksi kalimat-kalimat yang ada pada dokumen sumber [5]. Sedangkan metode peringkasan abstraktif bekerja dengan penyusunan kalimat yang baru, yang sesuai dengan aturan tata bahasa manusia. Metode abstraktif sangat sulit dilakukan secara otomatis, karena sistem harus mencocokkan tata bahasa yang tepat dengan model tata bahasa manusia [6].

Peringkasan teks otomatis modul pembelajaran berbahasa Indonesia menggunakan metode *Cross Latent Semantic Analysis* (CLSA) menghasilkan nilai rata-rata, *precision*, dan *recall* pada *compression rate* 20% berturut-turut yaitu 0.3853, 0.432, dan 0.3715 dan termasuk rendah [5]. Algoritma *textrank* untuk *automatic text summarization* pada dokumen berbahasa Indonesia, menyarankan agar melakukan proses penghilangan *stopword* agar kata-kata yang tidak mempunyai makna penting dalam tugas pengolahan teks, seperti kata hubung tidak dimasukkan dalam proses penghitungan *similarity* [1]. Berdasarkan literatur review oleh [4] menjelaskan bahwa sejauh ini peringkasan dokumen teks didominasi oleh teknik yang bersifat ekstraktif dan metode-metode *unsupervised*.

Metode lainnya yang telah diusulkan seperti *machine learning*, teruji mampu menghasilkan ringkasan yang baik, tetapi masih memerlukan banyak data latih. Metode berbasis *unsupervised* kebanyakan dikembangkan untuk peringkasan teks yang bersifat ekstraktif, salah satunya yang terdapat pada algoritma *textrank*. Metode kerja algoritma *textrank* ialah dengan cara memberikan skor kepada tiap kalimat, kemudian memilih kalimat dengan skor terbaik sebagai ringkasan akhir [4].

Keunggulan dari algoritma *textrank* yaitu tidak membutuhkan pengetahuan mendalam tentang tata bahasa dan tidak memerlukan data latih untuk dapat meringkas teks [7]. Pemilihan kalimat untuk menjadi ringkasan didasarkan pada formulasi matematis tertentu saja. Penelitian ini menyelidiki teknik pemilihan kandidat kalimat sebagai penyusun ringkasan, dengan menggunakan algoritma *textrank*. Asumsinya *textrank* bekerja akan memilih sejumlah kalimat yang dianggap berisi informasi penting dari dokumen dan menyusunnya dengan cara yang mirip dengan bagaimana manusia memilih kalimat di dalam dokumen yang panjang [8].

Penelitian yang dilakukan pada [9] menerapkan algoritma *textrank* untuk *multidocument summarization*, namun masih ditemukan kalimat serupa dalam hasil ringkasan.

Pada umumnya penelitian mengenai peringkasan teks otomatis mengambil ringkasan pada rangking teratas skor *textrank* saja, namun pada penelitian ini mencoba

menyelidiki pengambilan hasil ringkasan, tidak hanya bagian awal saja (*ranking* teratas), namun ringkasan dibagi menjadi 3 bagian, yakni ringkasan pada *ranking* awal, ringkasan pada *ranking* tengah, dan ringkasan pada *ranking* akhir yang diurutkan berdasarkan skor yang didapat dari algoritma *textrank* untuk setiap kalimat.

Hasil ringkasan yang dibentuk dari ketiga cara pemilihan kalimat ringkasan tersebut diharapkan lebih maksimal dalam menyampaikan fakta yang lebih banyak dan kaya informasi. Pertimbangannya, ringkasan yang berisi kalimat-kalimat yang mirip atau ide pokok yang diulang-ulang tentu memboroskan waktu membacanya. Hasil ringkasan tersebut dievaluasi menggunakan *ROUGE* (*Recall-Oriented Understudy for Gisting Evaluation*). *ROUGE* dihitung dengan membandingkan hasil ringkasan oleh sistem dengan ringkasan manual oleh manusia (*gold standard*).

Penerapan algoritma *textrank* pada peringkasan teks otomatis diharapkan dapat menghasilkan ringkasan yang lebih akurat sesuai dengan yang diharapkan. Penelitian ini bertujuan untuk menerapkan algoritma *textrank* dalam peringkasan artikel berbahasa Indonesia, menyelidiki cara pemilihan kandidat kalimat ringkasan berdasarkan posisi kalimat dari ranking *textrank*, dan kemudian mengukur relevansi hasil peringkasan menggunakan *ROUGE score* pada 300 dokumen berbahasa Indonesia.

2 LANDASAN TEORI

2.1 Peringkasan Teks Otomatis

Peringkasan teks adalah proses menghasilkan ringkasan dari suatu artikel atau dokumen yang sangat panjang oleh komputer. Peringkasan ekstraktif memilih beberapa kalimat dari dokumen asli untuk merepresentasikan dokumen secara keseluruhan tanpa perubahan dari struktur kalimat-kalimat tersebut. Sedangkan peringkasan abstraktif mengatur ulang kalimat-kalimat menjadi ringkasan berdasarkan kata-kata inti yang terdapat pada dokumen asli.

2.2 Text preprocessing

Text preprocessing merupakan langkah pertama dalam proses penambangan teks (*text mining*). Pada umumnya dokumen/artikel mempunyai bentuk yang tidak terstruktur. Oleh karena itu, diperlukan suatu proses yang dapat mengubah bentuk data yang sebelumnya tidak terstruktur ke dalam bentuk data yang terstruktur. Proses perubahan ini dikenal dengan istilah *text preprocessing* [10]. Tahapan-tahapan dalam *text preprocessing* yang dapat dilakukan antara lain adalah [11] sebagai berikut:

1. *Case Folding*: proses perubahan huruf seluruh teks menjadi huruf kecil (*lower case*).

2. *Tokenization*: proses pemecahan kalimat menjadi token-token (kata-kata tunggal) dengan spasi sebagai delimiter.
3. *Remove Punctuation*: proses penghapusan tanda baca, angka, dan simbol.
4. *Stemming*: proses mengubah kata-kata yang terdapat dalam suatu dokumen menjadi kata dasar (*root*).
5. *Stopword Removal*: penghapusan kata-kata umum yang sering muncul dan tidak memiliki arti seperti yang, di, ke, dari, dan dll.

2.5 Graf

Graf adalah sebuah struktur diskrit yang terdiri atas titik simpul (*vertex*) dan suatu himpunan pasangan tak berurutan yang menghubungkan dari titik-titik tersebut yang disebut sisi (*edge*) [6]. Graf dapat merepresentasikan hubungan antar kalimat di dalam suatu dokumen, berdasarkan kemiripan antar dokumen-dokumen tersebut,

2.6 ROUGE

ROUGE sendiri merupakan akronim dari *Recall-Oriented Understudy for Gisting Evaluation* atau dalam bahasa Indonesia berarti evaluasi intisari yang berorientasi *recall*. *ROUGE* merupakan standar pengujian peringkasan teks berbahasa Indonesia untuk mengukur kualitas hasil ringkasan dengan menghitung unit-unit token (*gram*) yang *overlap* seperti urutan kata dan pasangan-pasangan kata antara ringkasan manual yang dibuat oleh manusia sebagai referensi (*human annotator*) dan ringkasan yang *degenerate* oleh mesin. [3]. Ada beberapa macam jenis *ROUGE* yaitu *ROUGE-N*, *ROUGE-L* dan *ROUGE-SU* [4].

2.7 TextRank

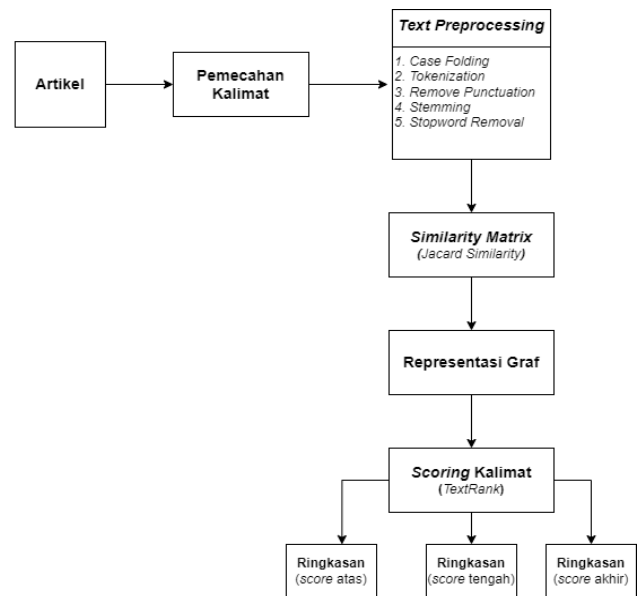
TextRank adalah salah satu algoritma yang berasal dari metode *pagerank*. *pagerank* adalah metode yang digunakan untuk memperkirakan seberapa penting sebuah node pada jaringan dan berbasis graf [12].

Algoritma perankingan berbasis graf adalah cara untuk menentukan seberapa penting sebuah *vertex* dalam graf, berdasarkan informasi yang diambil dari struktur graf keseluruhan secara rekursif. *textRank* mengekstraksi kalimat-kalimat beranking tinggi untuk dimasukkan ke dalam ringkasan.

3 METODOLOGI PENELITIAN

3.1 Tahapan Penelitian

Tahapan yang dilakukan dalam penelitian ini dimulai dengan pengumpulan dokumen, sebanyak 300 buah artikel dari berbagai topik melalui media online. Setiap artikel diproses dalam tahapan seperti pada Gambar 1.



Gambar 1. Tahapan Penelitian

3.1.2 Pemecahan Kalimat

Pemecahan kalimat membagi teks pada sebuah dokumen menjadi beberapa kalimat. Tanda pemisah (*delimiter*) yang dipakai pada proses *sentences splitting* adalah tanda titik (*.*), tanda tanya (*?*) dan tanda seru (*!*).

3.1.3 Text Preprocessing

Text preprocessing bertujuan untuk mengubah bentuk teks (kata atau kalimat) agar sesuai dengan kebutuhan *task*/tugas yang dikerjakan (problem yang akan dipecahkan). Tahapan-tahapan umum dalam *text preprocessing* meliputi *case folding*, *tokenization*, *remove punctuation*, *stemming*, dan *stopword removal*.

Tabel 1 di bawah ini menunjukkan contoh kalimat dari satu artikel di dalam korpus, yang telah melalui tahap *text preprocessing*, dengan judul artikel “Anak Di Bawah Umur Banyak Terlibat Tindak Kriminal Ini Kata Kapolres Pekanbaru” (sumber riaulink.com).

3.1.4 Pembentukan Similarity Matrix

Tahap pembentukan matriks kemiripan (*similarity matrix*) adalah metode untuk menghitung nilai kesamaan antar kalimat. Salah satu metode yang dapat digunakan untuk menghitung kemiripan kalimat adalah metode *jaccard similarity* [13].

Tabel 1. Hasil Preprocessing

Kalimat Awal	Hasil Preprocessing
1. PEKANBARU, RIAULINK.COM - Sangat disayangkan mulai banyaknya kasus pencurian dengan kekerasan (Curas) maupun pencurian sepeda motor (Curanmor) yang didalamnya ikut	1. pekanbaru riaulinkcom sayang banyak curi keras curas curi sepeda motor curanmor dalam libat anak anak umur perhati serius

<p>terlibat anak-anak di bawah umur harus menjadi perhatian serius.</p> <p>2. Dimana diketahui dari beberapa kali gelar press release kasus curas, maupun curanmor yang telah dilakukan pihak Polresta Pekanbaru beserta jajaran Polsek Polresta Pekanbaru masih ada mendapati anak di bawah umur terlibat melakukan tindak pidana tersebut.</p> <p>3. Kapolresta Pekanbaru Kombes Pol Nandang Mu'min Wijaya saat diwawancarai tanggapannya melihat masih adanya anak dibawah umur yang terlibat melakukan tindak pidana, mengatakan bahwa anak-anak tersebut melakukan tindak kejahatan karena salah pergaulan.</p> <p>4. "Kita mengimbau kepada para orangtua agar benar-benar bisa mengawasi anaknya di lingkungan pergaulan, karena salah pergaulan bisa membawa anak-anak menjadi pelaku kriminal, karena memang tantangan di kota besar ini banyak sekali, oleh karena itu perlu pengawasan orangtua, dari lingkungan rumah dulu yang paling utama," ucap Kapolresta, saat diwawancarai Kamis (16/1/2020).</p> <p>5. "Termasuk juga di lingkungan pendidikan, benar-benar di ajarkan pelajaran-pelajaran terkait dengan keagamaan maupun moralitas itu yang paling penting, agar anak-anak kita ini nanti bisa menjadi generasi harapan bangsa," tambah Kapolresta Pekanbaru.</p> <p>6. Lebih lanjut Kapolresta menyebutkan untuk meminimalisir angka kejahatan yang dilakukan anak-anak dibawah umur, pihak Polresta sudah melakukan upaya sosialisasi dengan cara program pembinaan.</p> <p>7. "Kita (Polresta Pekanbaru, red) sudah banyak program-program pembinaan, pembinaan ketertiban masyarakat yang dilaksanakan oleh jajaran satuan binmas maupun Polsek dan para bhabinkamtibmas untuk melakukan kegiatan pembinaan, penyuluhan dan sosialisasi terkait untuk mencegah kejahatan kejahatan baik kejahatan jalan, ataupun konvensional, narkoba maupun kekerasan didalam rumah tangga dan lainnya," jelasnya.</p> <p>8. Kegiatan yang paling penting itu adalah kegiatan yang bersifat bekerjasama dengan dunia pendidikan, disekolah-sekolah ya kita kerjasama untuk melakukan kegiatan penyuluhan kesekolah-</p>	<p>2. mana kali gelar press release curas curanmor polresta pekanbaru serta jajar polsek polresta pekanbaru dapat anak-anak terlibat tindak pidana</p> <p>3. kapolresta pekanbaru kombes pol nandang mumin wijaya wawancara tanggap anak-anak dibawah umur terlibat tindak pidana anak-anak tindak jahat salah gaul</p> <p>4. imbau orangtua awas anak lingkungan gaul salah gaul bawa anak-anak laku kriminal karena tantang kota awas orangtua lingkungan rumah utama kapolresta wawancara riaulinkcom Kamis 16/1/2020</p> <p>5. lingkungan didik ajar pelajaran-pelajaran kait agama moralitas anak-anak generasi harap bangsa kapolresta pekanbaru</p> <p>6. kapolresta meminimalisir angka jahat anak-anak dibawah umur polresta upaya sosialisasi program bina</p> <p>7. polresta pekanbaru red program program bina bina tertib masyarakat laksana jajar satu binmas polsek bhabinkamtibmas giat bina suluh sosialisai kait cegah jahat jahat jalan konvensional narkoba keras dalam rumah tangga</p> <p>8. giat giat sifat bekerjasama dunia didik disekolah-sekolah ya kerjasama giat suluh kesekolah-sekolah mengimbau ajar hindar narkoba hindar tindak kriminalitas</p> <p>9. singgung motif asing dengar laku kriminal tindak jahat karena main judi online kapolresta pekanbaru laksana razia rutin tingkat lokasi warnet</p> <p>10. laksana giat rutin tingkat jajar jajar satu fungsi pre-emptif preventif satu binmas satu sabhara jajar polsek bersamasama giat giat rutin tingkat dilokasilokasi warnet jam operasional lebih batas himbau ingat milik warnet patuh atur</p> <p>11. warnetwarnet disalahgunakan izin browsing disalahgunakan giat main game indikasi malam izin warnet tentu larut malam pagi giat</p>	<p>sekolah untuk mengimbau dan mengingatkan kepada pelajar untuk menghindari narkoba kemudian menghindari tindakan kriminalitas.</p> <p>9. Saat disinggung soal motif yang tidak asing didengar lagi dari para pelaku kriminal melakukan tindak kejahatan hanya karena untuk bermain judi online, Kapolresta Pekanbaru tegas mengatakan bahwa pihaknya saat ini sering melaksanakan razia rutin yang ditingkatkan dilokasi warnet.</p> <p>10. "Saat ini kita kan sering melaksanakan kegiatan rutin yang ditingkatkan baik jajaran-jajaran satuan fungsi pre-emptif preventif baik satuan binmas, satuan sabhara, termasuk juga jajaran polsek bersama-sama melakukan kegiatan-kegiatan rutin yang ditingkatkan, dilokasi-lokasi warnet yang jam operasionalnya melebihi batas, itu kita himbau dan ingatkan pada pemilik warnet untuk mematuhi aturan yang ada," katanya.</p> <p>11. "Jangan sampai warnet-warnet ini disalahgunakan yang seharusnya izinya untuk browsing, tapi disalahgunakan menjadi kegiatan permainan game, dan biasanya memang terindikasi lewat tengah malam, padahal sebenarnya izin warnet ditentukan tidak sampai lewat larut malam ataupun sampai pagi, dan kegiatan rutin yang ditingkatkan ini sering kita lakukan baik dari jajaran Polresta maupun jajaran Polsek," tutup Kapolresta</p> <p>rutin tingkat laku jajar polresta jajar polsek tutup kapolresta</p>
---	---	---

Metode *jaccard similarity* menghitung kesamaan antar 2 kalimat berdasarkan jumlah kata unik yang sama berbanding dengan total kata di kedua kalimat. Metode ini diambil dengan pertimbangan bahwa dalam satu artikel, hanya sedikit jumlah kata unik yang bisa dihasilkan, sehingga perhitungan menggunakan *jaccard similarity* dapat dilakukan dengan mudah dan efisien, dengan formula pada persamaan (1) berikut.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

Dimana $|A \cap B|$ adalah jumlah kata unik yang sama antara kalimat A dan B; $|A \cup B|$ adalah jumlah seluruh kata pada kalimat A dan B, dan $J(A,B)$ adalah nilai *jaccard similarity* yang akan dihitung pada kalimat A dan B.

Sebagai contoh, penerapan *jaccard similarity* untuk kalimat 2 dan kalimat 3 hasil *preprocessing* Tabel 1, dapat dihitung dengan cara sebagai berikut.

Kalimat 2:

“mana kali gelar press release curas curanmor polresta **pekanbaru** serta jajar polsek polresta **pekanbaru** dapat **anak umur libat tindak pidana**”

Kalimat 3:

“kapolresta **pekanbaru** kombes pol nandang mumin wijaya wawancara tanggap **anak** bawah **umur libat tindak pidana anak anak tindak jahat salah gaul**”

Bila kata yang tercetak tebal adalah kata yang sama (beririsan), ada di kedua kalimat, maka hasil perhitungan kemiripan antara kalimat 2 dan 3 menggunakan *jaccard similarity* adalah:

$$Sim(S_2, S_3) = 6/30 = 0.2$$

Tabel 2. Perhitungan Jaccard Similariry

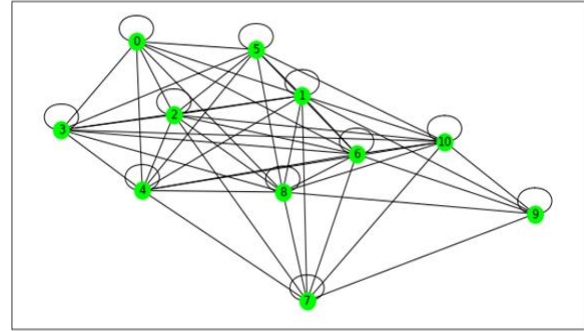
Sim	S1	S2	S3	S4	S5
S1	1	0.214	0.133	0.058	0.074
S2	0.214	1	0.2	0.027	0.068
S3	0.133	0.2	1	0.151	0.107
S4	0.058	0.027	0.151	1	0.1
S5	0.074	0.068	0.107	0.1	1
S6	0.076	0.111	0.2	0.066	0.086
S7	0.076	0.1	0.047	0.022	0.054
S8	0	0.03	0.03	0	0.074
S9	0.028	0.055	0.117	0.111	0.064
S10	0	0.048	0	0	0
S11	0	0.083	0.026	0.051	0.03

Sim	S6	S7	S8	S9	S10	S11
S1	0.076	0.076	0	0.028	0	0
S2	0.111	0.1	0.03	0.055	0.048	0.083
S3	0.2	0.047	0.03	0.117	0	0.026
S4	0.066	0.022	0	0.111	0	0.03
S5	0.086	0.054	0.074	0.064	0	0.03
S6	1	0.117	0	0.066	0	0.064
S7	0.117	1	0.05	0.069	0.133	0.093
S8	0	0.05	1	0.028	0.025	0.027
S9	0.066	0.069	0.028	1	0.09	0.171
S10	0	0.133	0.025	0.097	1	0.15
S11	0.064	0.093	0.027	0.171	0.15	1

3.1.5 Representasi Graf

Pada tahap pembentukan graf, *vertex* merepresentasikan kalimat-kalimat yang berada dalam dokumen sedangkan *edge* adalah representasi dari *similarity* antara dua kalimat yang saling berelasi. *Similarity* menyatakan hubungan antar *vertex* yang dihitung berdasarkan persamaan (1). Pada gambar 2.

adalah contoh representasi graf dari dokumen yang terdiri dari 11 kalimat (indeks kalimat dimulai dari nol).



Gambar 2. Representasi Graf berdasarkan similarity pada Tabel 2

3.1.6 Scoring Kalimat

Tahap *Scoring Kalimat* bertujuan mendapatkan skor akhir dari suatu *vertex*. Semakin besar skor akhir dari suatu *vertex*, maka semakin banyak informasi di dalamnya. Pada algoritma *textrank*, proses perangkingan dilakukan dengan menggunakan rumus untuk graf tidak berarah dan berbobot [1]. Kalimat yang sudah memperoleh skor, kemudian diberi peringkat berdasarkan skor tertinggi ke skor yang terendah. Berikut merupakan rumus tahap perangkingan kalimat menggunakan algoritma *textrank*

$$WS(S_i) = (1 - d) + d * \sum_{V_k \in Adj(V_j)} \frac{W_{ji}}{\sum_{V_k \in Adj(V_j)} W_{jk}} WS(S_j) \quad (2)$$

Dimana *WS* adalah *weight sentence* (skor kalimat); *d* adalah *dumping factor* yang diisi dengan 0.85; *V_i* adalah *vertex* yang akan dihitung skornya; *V_j* adalah *vertex* yang bertetangga dengan *V_i*; *V_k* adalah *vertex* yang bertetangga dengan *V_j*; *W_{ji}* adalah nilai *jaccard similarity* kalimat yang akan dihitung; *W_{jk}* adalah kalimat yang terhubung dengan *W_j*, *S_j* (bernilai = 1).

3.1.7 Pembentukan Ringkasan

Pembentukan ringkasan adalah proses penggabungan dari kalimat-kalimat yang terpilih sebagai ringkasan berdasarkan skor pada persamaan (2). Jumlah kalimat yang dipilih sesuai dengan *compression rate* (CR), yaitu seberapa padat ringkasan yang akan dihasilkan dibandingkan dokumen aslinya. Dalam penelitian ini digunakan *compression rate* 30% dan 50%.

Sebagai kontribusi penelitian, peneliti pencoba menyelidiki apakah hasil ringkasan akan lebih baik bila menggunakan kelompok kalimat dengan skor rendah, atau skor menengah, selain metode asli yang hanya mengambil *top-n* kalimat dengan score tinggi.

Kalimat-kalimat kandidat ringkasan, dikelompokkan menjadi tiga bagian, yaitu bagian awal, bagian tengah dan bagian akhir. Kalimat-kalimat yang

memperoleh ranking tertinggi pada proses pemeringkatan dikatego-rikan sebagai ringkasan awal (*top-n/skor tertinggi*), Kalimat-kalimat yang mendapat ranking pertengahan sebagai ringkasan tengah (*middle-n/skor menengah*), sedangkan kalimat-kalimat yang mendapat-kan ranking terendah sebagai ringkasan akhir (*bottom-n/skor teren-dah*). Pertimbangannya adalah, kalimat yang skornya tinggi akan memiliki kesamaan yang tinggi, sehingga kurang bermanfaat kalau dijadikan ringkasan karena informasinya akan relative sama. Sebaliknya, kalimat-kalimat dengan skor rendah, berarti kurang mirip bahkan berbeda, mungkin akan berguna bila terpilih sebagai ringkasan karena membawa informasi baru.

Berikut merupakan contoh pengelompokan skor kalimat algoritma *textrank* dengan *compression rate* 50%, dari 11 kalimat dokumen awal, dipilih 5 kalimat sebagai ringkasan.

Ranking ke-	Skor Kalimat	Nomor Kalimat
1	0.10258990411959588	3
2	0.0997119800434317	2
3	0.0954155363618634	9
4	0.09388190433304869	7
5	0.0931112015022097	6
6	0.0912895197435418	11
7	0.08946987602886734	5
8	0.08801099005183073	1
9	0.08576150784919453	4
10	0.08259628217578834	10
11	0.07816129779062853	8

Gambar 3. Pengelompokan Skor Algoritma *Textrank*

Kemudian kalimat-kalimat dalam masing-masing kelompok diurutkan berdasarkan ID kalimat, agar tampil berkesinambungan antara kalimat yang satu dengan kalimat yang lainnya di dalam ringkasan. Dengan demikian, ringkasan awal terdiri dari kalimat dengan ID {2,3,6,7,9}. Sedangkan ringkasan tengah terdiri dari kalimat {1,5,6,7,11}, dan ringkasan akhir terdiri dari kalimat {1,4,5,8,11}.

3.2 Evaluasi

Tahap evaluasi membandingkan hasil ringkasan oleh manusia (*gold-standard*) dengan hasil ringkasan oleh sistem, menggunakan metrik evaluasi *ROUGE*. *ROUGE* mengukur kualitas hasil ringkasan dengan menghitung jumlah kata yang beririsan antara ringkasan sistem dan ringkasan hasil yang menjadi referensi [3]. Pada penelitian ini mengambil nilai

precision, *recall*, dan pada *ROUGE-1*, *ROUGE-2* dan *ROUGE-L*, sebagaimana persamaan (3-9).

$$ROUGE - 1 recall = \frac{\text{jumlah unigram kata yang sama}}{\text{Total kata di ringkasan manual}} \quad (3)$$

$$ROUGE - 1 precision = \frac{\text{jumlah unigram kata yang sama}}{\text{Total kata di ringkasan sistem}} \quad (4)$$

$$ROUGE - 2 recall = \frac{\text{jumlah bigram kata yang sama}}{\text{Total kata di ringkasan manual}} \quad (5)$$

$$ROUGE - 2 precision = \frac{\text{jumlah bigram kata yang sama}}{\text{Total kata di ringkasan sistem}} \quad (6)$$

$$ROUGE - L recall = \frac{LCS(\text{sistem}, \text{manual})}{\text{Total kata di ringkasan manual}} \quad (7)$$

$$ROUGE - L precision = \frac{LCS(\text{sistem}, \text{manual})}{\text{Total kata di ringkasan sistem}} \quad (8)$$

$$f1 \text{ scores} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (9)$$

4 HASIL DAN PEMBAHASAN

4.1 Pemilihan Model *Textrank*

Dari dataset yang ada, dibagi menjadi 150 artikel sebagai *data training* untuk memilih model *textrank* yang terbaik di antara ketiga kelompok ringkasan (awal, tengah atau akhir). Model dipilih berdasarkan nilai *F1-score* tertinggi untuk *ROUGE-L* dari ringkasan 150 dokumen dibandingkan dengan *gold-standard*.

Model dipilih berdasarkan jumlah terbanyak dari kelompok ringkasan yang memperoleh nilai (persamaan 9) tertinggi. Kemudian model yang terpilih, diterapkan untuk *data testing*, yaitu sebanyak 150 artikel lainnya dalam dataset yang belum digunakan.

4.2 Model *Textrank* untuk CR 50%

Tabel 3 dibawah ini adalah cuplikan hasil ringkasan dengan *compression rate* 50% menggunakan 150 artikel, dengan data nilai *precision*, *recall*, dan dari *ROUGE-1*, *ROUGE-2*, dan *ROUGE-L*. Nilai maksimum dari tiap dokumen ditandai. Kemudian dilakukan pemilihan kelompok skor *textrank* berdasarkan nilai maksimal dari *ROUGE-L*.

Tabel 3. Data Evaluasi Ringkasan Model (CR=50%) pada *data training*

No	Judul Artikel	Kelompok Skor Textrank	ROUGE-1 (%)		
			R	P	F
1	7 Situs Berita Online...	Awal	53,66	64,23	58,47
		Tengah	53,05	56,86	54,89
		Akhir	70,12	66,09	68,05
2	Agen Rahasia Rusia ...	Awal	58,33	72,06	64,47
		Tengah	63,10	76,81	69,28
		Akhir	61,90	66,67	64,20
:	:	:	:	:	:
149	Satuan Reserse Kriminal ...	Awal	30,13	44,34	35,88
		Tengah	62,82	68,06	65,33
		Akhir	84,62	85,16	84,89
150	Seorang Wn Inggris Tersandung ...	Awal	52,90	71,30	60,74
		Tengah	65,81	76,12	70,59
		Akhir	69,68	74,48	72,00

No	Judul Artikel	Kelompok Skor Textrank	ROUGE-2 (%)		
			R	P	F
1	7 Situs Berita Online...	Awal	37,66	52,94	44,01
		Tengah	40,17	44,44	42,20
		Akhir	60,7	57,8	59,2
2	Agen Rahasia Rusia ...	Awal	47,17	64,10	54,35
		Tengah	53,77	67,86	60,00
		Akhir	52,83	59,67	56,00
:	:	:	:	:	:
149	Satuan Reserse Kriminal ...	Awal	20,20	34,19	25,40
		Tengah	55,56	63,22	59,14
		Akhir	83,33	83,76	83,54
150	Seorang Wn Inggris Tersandung ...	Awal	40,18	61,54	48,62
		Tengah	59,36	70,65	64,52
		Akhir	61,64	67,84	64,59

No	Judul Artikel	Kelompok Skor Textrank	ROUGE-L (%)			S max pada-
			R	P	F	
1	7 Situs Berita Online...	Awal	53,66	64,23	58,47	Akhir
		Tengah	52,44	56,21	54,26	
		Akhir	70,1	66,1	68,0	
2	Agen Rahasia Rusia ...	Awal	58,33	72,06	64,47	Tengah
		Tengah	61,90	75,36	67,97	
		Akhir	60,71	65,38	62,96	
:	:	:	:	:	:	
149	Satuan Reserse Kriminal ...	Awal	28,21	42,51	33,59	Akhir
		Tengah	60,90	65,97	63,33	
		Akhir	84,62	85,16	84,89	
150	Seorang Wn Inggris Tersandung ...	Awal	52,26	70,43	60,00	Akhir
		Tengah	65,16	75,37	69,90	
		Akhir	69,68	74,48	72,00	

Setelah itu dihitung total perolehan kelompok skor terhadap 150 dokumen tersebut. Kelompok yang terbanyak dipilih sebagai model metode *textrank* yang akan digunakan untuk pengujian terhadap *data testing*.

Tabel 4. Statistik Perolehan Kelompok Skor CR=50%

Posisi Kelompok Skor dgn tertinggi	Jumlah dokumen
Awal	7
Tengah	45
Akhir	98

Dari eksperimen yang dilakukan, terpilih kelompok skor akhir yang akan menjadi model untuk *deployment* (yang akan digunakan untuk membuat ringkasan pada data baru). Berdasarkan Tabel 4 diatas, maka total ringkasan dengan jumlah terbanyak adalah kelompok skor akhir (*bottom-n*) sebanyak 98 artikel, sehingga dipilih sebagai model untuk sistem *deployment* pada evaluasi ringkasan data testing.

4.3 Evaluasi Ringkasan pada Data Testing (50%)

Tabel 5 berikut merupakan data hasil evaluasi ringkasan uji dengan *compression rate* 50% menggunakan 150 artikel. Terlihat nilai untuk masing-masing *ROUGE-1*, *ROUGE-2* dan *ROUGE-L* mencapai di atas 60%, dengan rata-rata berturut-turut adalah 68.76%, 60.60% dan 68.29%.

4.4 Model *Textrank* untuk CR 30%

Tabel 6 berikut merupakan data pemilihan model dengan CR=30% menggunakan 150 artikel *data training*. Setelah itu dilakukan pemilihan skor *textrank* yang maksimal berdasarkan nilai dari *ROUGE-L*, berdasarkan posisi kelompok skor awal (*top-n*), menengah (*middle-n*) atau akhir (*bottom-n*).

Tabel 5. Hasil Evaluasi Ringkasan Data Testing (CR=50%)

No	Judul Artikel	ROUGE-1 (%)		
		R	P	F
1	(HUT) Harian Pekanbaru MX	81,67	72,27	76,88
2	5 Kondisi Kesehatan Gigi dan Mulut yang Menandakan Pikiran Sedang Stres	75,00	65,69	70,04
:	:	:	:	:
149	Yuk Kenali Produk Vegan	64,41	67,86	66,09
150	Yuk Mengenal Masjid Quba	80,47	60,59	69,13
RATA-RATA		72,32	66,15	68,76

No	Judul Artikel	ROUGE-2(%)		
		R	P	F
1	(HUT) Harian Pekanbaru MX	74,72	63,50	68,65
2	5 Kondisi Kesehatan Gigi dan Mulut yang Menandakan Pikiran Sedang Stres	67,95	58,89	63,10
⋮				
149	Yuk Kenali Produk Vegan	62,50	60,81	61,64
150	Yuk Mengenal Masjid Quba	72,35	51,46	60,15
RATA-RATA		65,08	57,29	60,60

No	Judul Artikel	ROUGE-L(%)		
		R	P	F
1	(HUT) Harian Pekanbaru MX	81,33	71,98	76,37
2	5 Kondisi Kesehatan Gigi dan Mulut yang Menandakan Pikiran Sedang Stres	74,17	64,96	69,26
⋮				
149	Yuk Kenali Produk Vegan	62,71	66,07	64,35
150	Yuk Mengenal Masjid Quba	80,47	60,59	69,13
RATA-RATA		71,81	65,72	68,29

Tabel 6. Data Evaluasi Ringkasan Model (CR=30%) pada data training

No	Judul Artikel	Kelompok Skor Textrank	ROUGE-1(%)		
			R	P	F
1	7 Situs Berita Online...	Awal	29,41	40,70	34,15
		Tengah	34,45	38,68	36,44
		Akhir	31,93	41,30	36,02
2	Agen Rahasia Rusia ...	Awal	14,00	17,50	15,56
		Tengah	58,00	54,72	56,31
		Akhir	36,00	34,62	35,29
⋮					
149	Satuan Reserse Kriminal ...	Awal	8,51	16,67	11,27
		Tengah	73,40	69,00	71,13
		Akhir	44,68	42,42	43,52
150	Seorang Wn Inggris Tersandung ...	Awal	18,10	27,94	21,97
		Tengah	64,76	72,34	68,34
		Akhir	43,14	43,14	43,14

No	Judul Artikel	Kelompok Skor Textrank	ROUGE-2(%)		
			R	P	F
1	7 Situs Berita Online...	Awal	14,74	24,21	18,33
		Tengah	19,23	21,28	20,20
		Akhir	18,59	23,97	20,94
2	Agen Rahasia Rusia ...	Awal	1,00	1,00	1,00
		Tengah	42,28	44,44	46,28
		Akhir	18,97	20,00	19,47
⋮					
149	Satuan Reserse Kriminal ...	Awal	5,26	11,32	7,19
		Tengah	67,54	68,14	67,84
		Akhir	31,58	32,14	31,86
150	Seorang Wn Inggris Tersandung ...	Awal	4,48	8,22	5,80
		Tengah	53,73	62,07	57,60
		Akhir	26,79	28,30	27,52

No	Judul Artikel	Kelompok Skor Textrank	ROUGE-L(%)			S max pada-
			R	P	F	
1	7 Situs Berita Online...	Awal	28,57	39,53	33,17	Akhir
		Tengah	32,77	36,79	34,67	
		Akhir	31,09	40,22	35,07	
2	Agen Rahasia Rusia ...	Awal	12,00	15,00	13,33	Tengah
		Tengah	58,00	54,72	56,31	
		Akhir	34,00	32,69	33,33	
⋮						
149	Satuan Reserse Kriminal ...	Awal	8,51	16,67	11,27	Tengah
		Tengah	73,40	69,00	71,13	
		Akhir	43,62	41,41	42,49	
150	Seorang Wn Inggris Tersandung ...	Awal	17,14	26,47	20,81	Tengah
		Tengah	61,90	69,15	65,33	
		Akhir	41,18	41,18	41,18	

Tabel 7. Statistik perolehan kelompok skor CR=30%

Posisi Kelompok Skor dgn tertinggi	Jumlah dokumen
Awal	38
Tengah	65
Akhir	47

Tabel 7. diatas merupakan total hasil dari penjumlahan posisi kelompok skor *textrank* dengan tertinggi. Model dengan CR=30%, yaitu kelompok skor tengah sebanyak 65 artikel, digunakan sebagai evaluasi

ringkasan pada data *testing* untuk pengujian akhir sebagai model *deployment*.

4.5 Evaluasi Ringkasan pada Data Testing (30%)

Tabel 8 berikut adalah data hasil eksperimen menggunakan model terpilih CR=30%, menggunakan model skor tengah. Secara umum nilai rata-rata untuk *ROUGE-1*, *ROUGE-2* dan *ROUGE-L* menurun dibandingkan dengan CR=50%. Hal ini karena jumlah kalimat yang dibangkitkan sebagai hasil ringkasan lebih sedikit. Dari hasil yang diperoleh, metode ini sudah cukup baik karena dapat menghasilkan angka h seperti rata-rata penelitian serupa, yaitu di kisaran angka 30-50% untuk CR yang rendah (misalnya 30%). Dan nilai pada *ROUGE-1* dan *ROUGE-L* dapat melebihi 40%.

Tabel 8. Hasil Evaluasi Ringkasan Data Testing (CR=30%)

No	Judul Artikel	ROUGE-1 (%)		
		R	P	F
1	(HUT) Harian Pekanbaru MX	34,44	56,36	42,76
2	5 Kondisi Kesehatan Gigi dan Mulut yang Menandakan Pikiran Sedang Stres	53,42	51,32	52,35
⋮	⋮	⋮	⋮	⋮
149	Yuk Kenali Produk Vegan	53,42	51,32	52,35
150	Yuk Mengenal Masjid Quba	53,42	51,32	52,35
RATA-RATA		43,28	48,02	45,00

No	Judul Artikel	ROUGE-2 (%)		
		R	P	F
1	(HUT) Harian Pekanbaru MX	24,04	39,68	29,94
2	5 Kondisi Kesehatan Gigi dan Mulut yang Menandakan Pikiran Sedang Stres	42,68	44,32	43,58
⋮	⋮	⋮	⋮	⋮
149	Yuk Kenali Produk Vegan	42,68	44,32	43,58
150	Yuk Mengenal Masjid Quba	42,68	44,32	43,58
RATA-RATA		30,24	34,05	31,62

No	Judul Artikel	ROUGE-L (%)		
		R	P	F
1	(HUT) Harian Pekanbaru MX	34,44	56,36	42,76
2	5 Kondisi Kesehatan Gigi dan Mulut yang Menandakan Pikiran Sedang Stres	52,05	50,00	51,01
⋮	⋮	⋮	⋮	⋮
149	Yuk Kenali Produk Vegan	52,05	50,00	51,01
150	Yuk Mengenal Masjid Quba	52,05	50,00	51,01
RATA-RATA		41,91	46,52	43,59

5 SIMPULAN

Berdasarkan hasil yang diperoleh dari algoritma *textrank* pada dokumen Bahasa Indonesia di atas, pemilihan skor kalimat tidak harus mengambil *top-n* saja. Dari pengujian yang sudah dilakukan, untuk CR=50%, terlihat bahwa hasil ringkasan lebih baik bila mengambil kelompok kalimat dengan skor *textrank* di posisi akhir (*bottom-n*), dengan rata-rata *s* 68,76% pada *ROUGE-1*, 60,60% pada *ROUGE-2*, dan 68,29% pada *ROUGE-L*.

Sedangkan untuk CR=30%, hasil ringkasan lebih baik dengan mengambil kelompok skor *textrank* bagian tengah (*middle-n*), dengan rata-rata *s* 45,00%, 31,62% dan 43,59% pada *ROUGE-1*, *ROUGE-2*, dan *ROUGE-L* secara berturut-turut.

Didapatkan kesimpulan bahwa peringkasan teks otomatis dengan *jaccard similarity* pada metode *textrank*, kurang baik bila mengambil kelompok skor awal, disebabkan kalimat-kalimat yang mirip lebih banyak terpilih sebagai ringkasan. Hal ini kurang dapat memberikan lebih banyak fakta dari artikel yang diringkaskan.

KEPUSTAKAAN

- [1] J. Pragantha, T. Informatika, F. T. Informasi, and U. Tarumanagara, "Penerapan Algoritma Textrank Untuk Automatic Summarization Pada Dokumen Berbahasa Indonesia," vol. 1, no. 1, pp. 71–78, 2017.
- [2] J. Kurniawati and S. Baroroh, "Literasi Media Digital Mahasiswa Universitas Muhammadiyah Bengkulu," *J. Komun.*, vol. 8, no. 2, pp. 51–66, 2016, [Online]. Available: <https://journal.umy.ac.id/index.php/jkm/article/download/2069/2586>
- [3] D. P. Ismi and F. Ardianto, "Peringkasan Ekstraktif Teks Bahasa Indonesia dengan Pendekatan Unsupervised Menggunakan Metode Clustering," *Cybernetics*, vol. 3, no. 02,

- p. 90, 2020, doi: 10.29406/cbn.v3i02.2290.
- [4] Y. Yuliska and K. U. Syaliman, "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, doi: 10.25299/itjrd.2020.vol5(1).4688.
- [5] Y. M. Sari and N. S. Fatonah, "Peringkasan Teks Otomatis pada Modul Pembelajaran Berbahasa Indonesia Menggunakan Cross Latent Semantic Analysis (CLSA)," vol. 7, no. 2, pp. 153–159, 2021.
- [6] F. Kiyoumars, "Evaluation of Automatic Text Summarizations based on Human Summaries," *Procedia - Soc. Behav. Sci.*, vol. 192, pp. 83–91, 2015, doi: 10.1016/j.sbspro.2015.06.013.
- [7] M. P. Akademik and M. Pematangsiantar, "Pengenalan algortima dalam model peringkasan teks untuk mempercepat pekerjaan akademik," vol. 1, no. 2, pp. 19–26, 2021.
- [8] M. A. Zamzam, "Sistem Automatic Text Summarization Menggunakan Algoritma Textrank," *Matics*, vol. 12, no. 2, pp. 111–116, 2020, doi: 10.18860/mat.v12i2.8372.
- [9] D. Gunawan, S. H. Harahap, and R. F. Rahmat, "Peringkasan multi-dokumen dengan menggunakan Textrank dan Relevansi Marjinal Maksimal untuk Teks di Bahasa Indonesia," 2019.
- [10] F. S. Jumeilah, "Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 1, pp. 19–25, 2017, doi: 10.29207/resti.v1i1.11.
- [11] M. Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J. & Nithya, "Preprocessing Techniques for Text Mining Preprocessing Techniques for Text Mining," *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. October 2014, pp. 7–16, 2015.
- [12] M. F. Fakhrezi, M. A. Bijaksana, and A. F. Huda, "Implementation of Automatic Text Summarization with Textrank Method in the Development of Al-Qur'an Vocabulary Encyclopedia," *Procedia Comput. Sci.*, vol. 179, no. 2020, pp. 391–398, 2021, doi: 10.1016/j.procs.2021.01.021.
- [13] A. F. Zulfikar, "Pengembangan Algoritma Stemming Bahasa Indonesia dengan Pendekatan Dictionary Base Stemming untuk Menentukan Kata Dasar dari Kata yang Berimbuhan," *J. Inform. Univ. Pamulang*, vol. 2, no. 3, p. 143, 2017, doi: 10.32493/informatika.v2i3.1443.