PERINGKAS TEKS OTOMATIS DOKUEM TUNGGAL DAN MULTI BAHASA MENGGUNAKAN METODE TF-IDF

Aa Zezen Zenal Abidin [*1], Eka Yuda Rayi A [#2]

Program Studi Teknik Informatika, STMIK Subang
Jl. Marsinu No. 5 - Subang, Tlp. 0206-417853 Fax. 0206-411873
E-mail: zezen2008@yahoo.com [*1], beat.collector27@gmail.com [#2]

Abstrak

Sistem peringkas teks otomatis berita kesehatan multi bahasa bisa digunakan oleh pembaca untuk meringkas teks berita kesehatan beserta terjemahanya untuk kebutuhan berbagai jenis aktivitas. Untuk mengembangkan sebuah perangkat lunak yang sesuai dan mudah dipahami untuk user dalam membaca berita dan menterjemahkannya pada bahasa international. Berita kesehatan dilakukan proses *teks processing* meliputi hapus tanda baca, stopword, *stemming*, tokenizing, pembobotan kata dan pembobotan kalimat. Setelah *teks processing* setiap kalimat akan mempunyai bobotnya masing masing dari yang terkecil hingga terendah. K8 mendapatkan bobot 37.19723289, K9 mendapatkan bobot 17.89999416, K10 mendapatkan bobot 16.52464106 dan K3 mendapatkan bobot 14.77709596. metode yang digunakan adalah Metode Term frequency inverse document frequency (TF-IDF). Sistem menerima entri berita, translate, dokumen menjadi kalimat, membuang karakter, memecahkanya menjadi kata, memberikan nilai bobot pada kata, menjumlahkan nilai bobot, menghitung nilai idf dan TD-IDF sehingga bisa didapatkan nilai bobot dari setiap kalimat yang akan menghasilkan peringkat paling tertinggi. Bahasa pemrograman yang digunakan yaitu PHP dan DBMS menggunakan MySQL. Editor menggunakan Sublime dan Tools menggunakan Xampp 3.2.2 Sistem peringkas teks berita kesehatan otomatis multi bahasa pembaca tidak perlu menyita waktu yang cukup lama untuk membaca dan menerjemahkan. Sistem peringkas teks berita kesehatan dengan melalui tahap pengujian pembobotan kalimat manual dan otomatis dengan hasil nilai yang sama dan sistem ini dapat mengetahui isi penting dari berita yang diinputkan, dengan memiliki akurasi uji responden 54,17 %.

Kata kunci: Teks mining, TF-IDF, peringkas otomatis, berita kesehatan.

1. Pendahuluan

Menurut riset para ahli minat baca masyarakat Indonesia hanya mencapai 0,01% buku per tahunya sementara dibandingkan dengan masyarakat Eropa atau Amerika mencapai 25-27% buku per tahunnya.Indonesia menempati urutan ke 39 dari 80 negara dalam kategori penguasaan Bahasa Inggris pada skala international. Fakta tersebut diperkuat berdasarkan survei dari **English Profiency Indeks (EE EPI),** bahwa Indonesia memperoleh nilai total 52,15 dalam mengukur kemampuan bahasa inggris negara-negara di dunia di jadikan sebagai patokan international untuk kemampuan bahasa inggris tingkat dewasa. Dari hasil survei yang telah dilakukan menunjukkan peringkat tiga besar dari tiga negara besar di asia, yaitu Singapura, Malaysia, dan Philipina, sementara Indonesia masih masuk pada level Menengah.

Berkembangnya teknologi internet berdampak pada bertambahnya juga jumlah situs berita berbahasa nasional maupun international dan menciptakan ledakan informasi. Bagi pembaca yang hanya memiliki sedikit waktu diperkirakan mereka tidak sempat membaca semua berita pada hari tersebut. Hal ini dikarenakan karena jumlahnya yang terlampau banyak. Selain itu, dengan data bahwa Indonesia berada pada level menengah terkait masalah pemahaman Bahasa Internasional, dengan banyaknya berita berbahasa international masyarakat Indonesia akan mengalami kesulitan untuk memahami isi berita.

peringkas teks otomatis dapat mengurangi waktu dalam membaca dokumen yang panjang dengan cara mengekstrak informasi yang paling penting [1]. Pada zaman sekarang peringkas

teks merupakan sebuah alat yang sangat dibutuhkan untuk pembaca. Situs-situs berita yang menyediakan informasi tiap detiknya selalu memperbarui berita. salah satu contohnya adalah berita kesehatan, karena berita kesehatan merupakan salah satu kategori yang paling banyak di baca daripada kategori yang lainya.Ringkasan sangat dibutuhkan untuk mendapatkan berita secara ringkas. Konsep sederhana ringkasan adalah mengambil bagian penting dari keseluruhan isi dari artikel.

Rendahnya minat intensitas membaca, rendahnya penguasaan bahasa Inggris dalam media mendorong melatarbelakangi penulis untuk memngambangkan suatu perangkat lunak peringkas teks dan menterjemahkannya dalam bahasa asing khusunya bahasa Inggris dimungkinkan dapat dihadirkankan dengan m perangkat lunak peringkas teks dokumen tunggal multi dokumen.

2. Tinjauan Pustaka

Peringkas Teks Otomatis adalah pembuatan bentuk yang lebih singkat dari suatu teks dengan memanfaatkan sistem yang dijalankan dan dioperasikan pada komputer[2]. Dua diantaranya peringkas teks dalam bidang medis adalah seperti penelitian oleh [3][1].Banyak Teknik yang digunakan dalam peringkasan ini, antara lain Teknik Pendekatan statistika dan Teknik Pendekatan dengan Naturan Language Analysis. Beberapa Teknik Pendekatan statistika adalah Sebagai berikut [2]:

a) Teknik *Word Frequency*
b) *Position in text*
c) *Cue Words and heading*
d) *Setence position*

Teknik Pendekatan dengan Naturan Language Analysis

a) *Inverse Term Frequency and NLP Technique*
b) *Lexical chain*,
c) *Maximal Marginal Relevance*

Terdapat dua pendekatan pada peringkas teks, yaitu ekstraksi (*shallower approaches*) dan abstraksi (*deeper approaches*). Pada bentuk ekstraksi, sistem menyalin unit-unit teks yang dianggap paling penting atau paling informatif dari teks sumber menjadi ringkasan. Unit-unit teks yang disalin dapat berupa klausa utama, kalimat utama, atau paragraf utama. Sedangkan teknik abstraksi melibatkan parafrase dari teks sumber, Teknik abstraksi mengambil intisari dari teks sumber, kemudian membuat ringkasan dengan menciptakan kalimat-kalimat baru yang mempersentasikan intisari teks sumber dalam bentuk berbeda dengan kalimat-kalimat pada teks sumber. Pada umumnya, abstraksi dapat meringkas teks lebih kuat daripaa ekstraksi, tetapi sistemnya lebuh sulit dikembangkan karena mengaplikasikan teknologi *natural language generation* maupun *natural language question* [4] yang merupakan bahasan yang dikembangkan tersendiri.

Berdasarkan jumlah sumbernya, sebuah ringkasan dapat dihasilkan dari satu sumber (single- document) atau dari banyak sumber (*multi-document*)[5]. Peringkasan single-document masukanya berupa sebuah teks dan keluaranya berupa sebuah teks baru yang lebih singkat. Pada peringkasan multi-document, masukan adalah beberapa dokumen teks yang memiliki tema sam, biasanya sudah ada dalam satu klaster kemudian akan dihasilkan keluaran berupa sebuah teks yang lebih singkat yan merangkum informasi-informasi utama pada klaster masukan.

Dalam metode ini pembobotan kata dalam sebuah dokumen dilakukan dengan mengalikan nilai Tf dan IDF. Pembobotan diperoleh berdasarkan jumlah kemunculan term dalam kalimat (TF) dan jumlah kemunculan term pada seluruh kalimat dalam dokum (IDF). Bobot suatu istilah semakin besar jika istilah tersebut sering muncul dalam suatu dokumen dan semakin kecil jika istilah tersenut muncul dalam banyak dokumen, Nilai IDF sebuah term dihitung menggunakan persamaan dibawah [2]:

Menghitung bobot (W) masing-masing dokumen dengan persamaan dibawah ini:

$$df=d1+d2+...+dn \quad \textbf{(1)}$$

Keterangan pada rumus 2.1

- Df : Total kata pada term
- D : frekuensi kata pada kalimat

$$id\ f = log\ (n/Df) \quad \textbf{(2)}$$

Keterangan pada rumus 2.2

- N : total dokumen
- Df : banyak dokumen yang mengandung kata yang dicari

$$W\ (t, d) = tf\ (t, d) * idf \quad \textbf{(3)}$$

Keterangan pada rumus 2.3

- d : dokumen ke – d
- t : kata ke -t dari kata kunci
- tf : banyaknya kata yang dicari pada sebuah dokumen
- IDF : Inversed Document Frequency

Kemudian baru melakukan proses pengurutan (sorting) nilai kumulatif dari W untuk setiap kalimat. Tiga kalimat dengan nilai W tersebesar dijadikan sebagai hasil dari ringkasan atau output dari peringkasan teks otomatis.

3. Analisa dan Pembahasan

Dilakukan pemecahan kaliamt sehingga dokumen artikel kesehatan sebagai kasus dalam penenlitian ini terbagi menjadi 13 kalimat seperti diperlihatkan dalam Tabel 1. setelah dilakukan proses tekenizing, yaitu penghilangan tanda bacaca,kemudian dilakukam proses stemming, dimana setip kalimat dipecah lagi menjadi daftar kata, dimana potongan tablenya diperlihatkan dalam Tabel 2.

Tabel 1. *Daftar Kalimat*

No.	Kalimat
K1	Jakarta, CNN Indonesia -- Puasa bukan hanya sekadar menahan lapar dan haus.
K2	Selain untuk menjalankan ibadah, puasa juga memberikan banyak manfaat untuk kesehatan..
K3	Penelitian terbaru yang dipublikasikan di Jurnal Cell Stem Cell menyatakan puasa dapat meningkatkan kemampuan sel batang usus untuk melakukan regenerasi secara signifikan.
K4	Sel batang usus merupakan sel yang berfungsi memelihara lapisan usus dan dapat memperbarui diri setiap lima hari.
K5	Sel batang berperan penting untuk memperbaiki setiap kerusakan saat luka atau infeksi terjadi..
K6	Namun saat orang bertambah usia, kemampuan regenerasi sel batang usus terus menurun, sehingga usus butuh waktu lama untuk pulih.
K7	Saat berpuasa, sel di usus ini memisahkan asam lemak yang dapat merangsang sel batang untuk meregenerasi.
K8	Para peneliti yang merupakan ahli biologi itu juga menemukan regenerasi itu berpotensi membantu orang yang memiliki infeksi pencernaan atau pasien kanker yang menjalani kemoterapi pulih lebih cepat.
K9	"Puasa memiliki banyak efek pada usus, termasuk mendorong regenerasi serta berpotensi digunakan (sebagai metode) pada setiap jenis penyakit yang terjadi pada usus, seperti infeksi atau kanker," kata salah seorang penulis studi yang merupakan Asisten Profesor Biologi di Massachusetts Institute of Technology (MIT) Omer Yilmaz dikutip dari *Antara*.
K10	Menurut Yilmaz, puasa berpeluang menjadi salah satu metode pengobatan untuk

No.	Kalimat
	meningkatkan jaringan ketahanan patologi yang menurun seiring berjalannya usia.
K11	Yilmaz dan beberapa ahli biologi melakukan penelitian ini pada tikus yang berpuasa selama 24 jam.
K12	Para peneliti lalu mengangkat sel batang usus dan mengembangkannya pada cawan petri.
K13	Peneliti menemukan kemampuan regenerasi sel batang usus pada tikus yang berpuasa meningkat dua kali lipat. (chs/chs).

Tabel 2. *Potongan Frekuenasi kata*

No	Term	Kata Per kalimat											
		K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	K11	K12
1	Jakarta	1	0	0	0	0	0	0	0	0	0	0	0
2	Cnn	1	0	0	0	0	0	0	0	0	0	0	0
3	Indonesia	1	0	0	0	0	0	0	0	0	0	0	0
4	Puasa	1	1	1	0	0	0	0	0	1	1	0	0
5	Puasa	0	0	0	0	0	0	1	0	0	0	1	0
6	Sekadar	1	0	0	0	0	0	0	0	0	0	0	0
7	Tahan	1	0	0	0	0	0	0	0	0	0	0	0
8	Lapar	1	0	0	0	0	0	0	0	0	0	0	0
9	Haus	1	0	0	0	0	0	0	0	0	0	0	0
10	Selain	0	1	0	0	0	0	0	0	0	0	0	0

Berikut ini adalah perhitungan nilai DF pada term tiap kalimat dapat di lihat sebagai berikut, hasil perhitungan DF diperlihatkan pada Tabel 3. Nilai IDF diperoleh dari Tabel 3, hasil IDF diperlihatkan dalam Tabel 4.

1) Perhitungan df term jakarta = tf (K1 + K2 + K3 + … + K13)
= 1 + 0 + 1 + …+ 0
= 1
2) Perhitungan df term cnn = tf (K1 + K2 + K3 + …+ K13)
= 1 + 0 + 0 + … + 0
= 1
3) Perhitungan df term indonesia = tf (K1 + K2 + K3 + … + K13)
= 1 + 0 + 0 + …+ 0
= 1

Tabel 3. Nilai DF

NO	Term	Df
1	Jakarta	1
2	Cnn	1
3	Indoensia	1
….		
134	Lipat	1

Pada Tabel 4 merupakan proses perhitungan nilai IDF, perhitungan dilakukan menggunakan Rumus 2, sebagai berikut:

1) Perhitungan idf term jakarta = log (N/df)

= log (13/1)
= 1.113943352

2) Perhitungan idf term cnn = log (N/df)
= log (13/1)
= 1.113943352

…

134) Perhitungan idf term lipat = log (N/df)
= log (13/1)
= 1.113943352

Tabel 4. Nilai IDF

NO	Term	Idf
		Log (n/df)
1	Jakarta	1.113943352
2	Cnn	1.113943352
3	indonesia	1.113943352
…		
134	Lipat	1.113943352

Penentuan nilai bobot kata dalam TF-IDF menggunakan Rumus 3, sebagai berkut:

K1 Perhitungan Wdt term jakarta = tf * idf
= 1 * 1.113943352
= 1.113943352

K2 Perhitungan Wdt term jakarta = tf* idf
= 0 * 1.113943352
= 0

…

K13 Wdt term jakarta = tf * idf
= 0 * 1.113943352
= 0

Hasil perhitungan nilai bobot disimpan dalam Tabel 5. Kumulatif seluruh kata yang terkandung dalam setiap kalimat diperlihatkan dalam Tabel 6. Kalimat terpilih diambil 4 kalimat sebagai rangking teratas, seperti diperlihatkan pada Tabel 7. demikian juag untuk ringkasan 30 %, diperoleh 4 kalimat terpilih sebagai ringkasan, seperti pada Tabel 8.

Tabel 5. Bobot kata

No	Term	K1	K2	K3	K4	K5
1	jakarta	1.113943352	0	0	0	0
2	Cnn	1.113943352	0	0	0	0
3	indonesia	1.113943352	0	0	0	0
…						
134	Lipat	0	0	0	0	0

Tabel 6. Urutan nilai bobot kalimat

Kalimat	Bobot
K9	33,197
K8	17,900
K10	16,525
K3	14,777

Kalimat	Bobot
K6	13,663
K4	10,524
K11	9,627
K5	9,285
K13	9,264
K7	8,447

Tabel 7. Nilai bobot kalimat terpilih sebagai ringkasan

Kalimat	Bobot
K9	33,197
K8	17,900
K10	16,525
K3	14,77

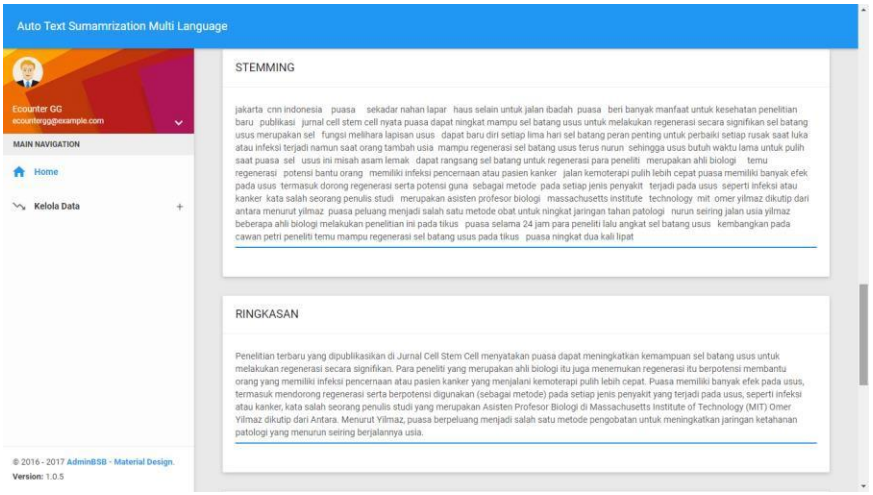
Tabel 8 Urutan kalimat hasil ringkasan 30 %

Kalimat	Bobot
K9	33,197
K8	17,900
K10	16,525
K3	14,77

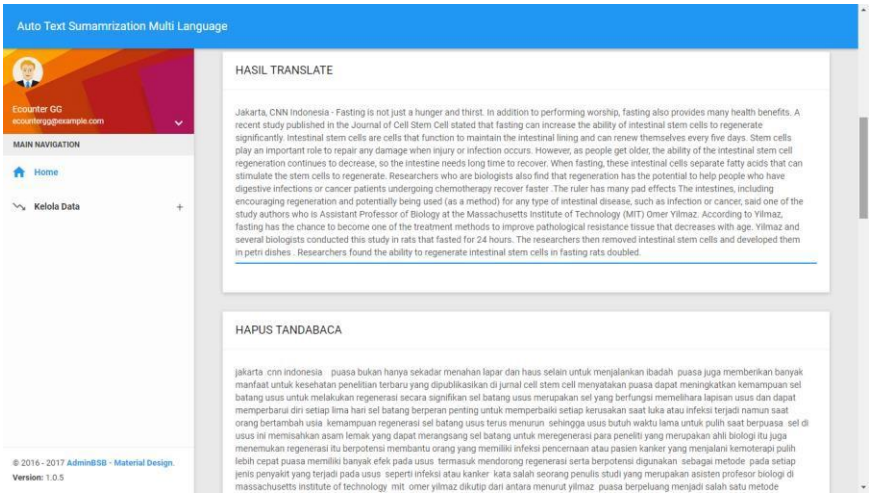
Tabel 9. Hasil Resume berdasarkan urutan bobot

No. K	Kalimat
K9	"Puasa memiliki banyak efek pada usus, termasuk mendorong regenerasi serta berpotensi digunakan (sebagai metode) pada setiap jenis penyakit yang terjadi pada usus, seperti infeksi atau kanker," kata salah seorang penulis studi yang merupakan Asisten Profesor Biologi di Massachusetts Institute of Technology (MIT) Omer Yilmaz dikutip dari *Antara*.
K8	Para peneliti yang merupakan ahli biologi itu juga menemukan regenerasi itu berpotensi membantu orang yang memiliki infeksi pencernaan atau pasien kanker yang menjalani kemoterapi pulih lebih cepat.
K10	Menurut Yilmaz, puasa berpeluang menjadi salah satu metode pengobatan untuk meningkatkan jaringan ketahanan patologi yang menurun seiring berjalannya usia.
K3	Penelitian terbaru yang dipublikasikan di Jurnal Cell Stem Cell menyatakan puasa dapat meningkatkan kemampuan sel batang usus untuk melakukan regenerasi secara signifikan.

Pememrosesan *translate* yang diolah hanya dokumen utuh bukan dokumen yang sudah melakukan tahap teks processing, hasil peringkasan teks diperlihatkan dalam Gambar 1, hasil translate diperlihatkan dalam Gambar 2. *Translator* tersebut menggunakan layanan Google Translate API secara otomatis akan memperoleh terjemahan yang dikelola oleh layanan tersebut. Terdapat 1 (satu) dokumen yang telah ditranslate oleh admin dan disimpan sebagai berikut:



Gambar 1. Tampilan ringkasan



Gambar 2. Hasil Translate

4. Hasil

Dokumen berita kesehatan dipecah menjadi kalaimat, dalam penelitian ini diperoleh 13 kalimat. Tiga belas kalimat yang tersedia dilakukan proses preprocessing dan penghitungan frekuensi kata, bobot kata dan akhirnya bobot kalimat. Ditentukan sebanyak 3 kalimat dengan bobot terbesar seperti pada Tabel 9, sebagai ringkasan. Ringkasan seeperti pada Gambar 1, kemudian di translate menggunakan Google API yang tersedia pada layanan google seperti diperlihatkan pada Gambar 2 sehingga menjadi system peringkas teks tunggal multi dokumen yaitu ringkasan dalam bahasa Indonesia dan ringkasan dalam bahasa Inggris. Kemiripan pada dokumen ini menggunakan tingkat persentase, dimana setiap dokumen akan dibandingkan dan dihitung kemudia dirubah dalam bentuk persen (%). Dengan mengambil 30 responden sebagain data sample untuk membandingkan tingkat akurasi ringkasan dari responden. Seperti pada Table 10.

Tabel 10 Perhitungan akurasi hasil ringkasan responden

No.	Jumlah Kalimat	Hasil Ringkasan Sistem	Akurasi Sistem	Hasil Ringkasan Reponden		Akurasi Responden
				Sesuai	Tidak Sesuai	
1	13	4	100	2	2	50
2	13	4	100	3	1	75
3	13	4	100	2	2	50
4	13	4	100	2	2	50
5	13	4	100	2	2	50
6	13	4	100	3	1	75
7	13	4	100	2	2	50
8	13	4	100	2	2	50
9	13	4	100	2	2	50
10	13	4	100	3	1	75
11	13	4	100	1	3	25
12	13	4	100	2	2	50
13	13	4	100	4	0	100
14	13	4	100	2	2	50
15	13	4	100	3	1	75
16	13	4	100	2	2	50
17	13	4	100	2	2	50
18	13	4	100	2	2	50
19	13	4	100	2	2	50
20	13	4	100	3	1	75
21	13	4	100	2	2	50
22	13	4	100	1	3	25
23	13	4	100	2	2	50
24	13	4	100	2	2	50
25	13	4	100	2	2	50
26	13	4	100	2	2	50
27	13	4	100	3	1	75
28	13	4	100	1	3	25
29	13	4	100	2	2	50
30	13	4	100	2	2	50
Total			3000	65	55	1625

Berdasarkan data yang ditampilkan pada Tabel 4.6 diatas untuk menentukan tingkat akurasi dilakukan dengan rumus:

$$Tingkat\ akurasi = \frac{Total\ akurasi\ responden}{Total\ akurasi\ sistem}\ x\ 100$$

Maka,

$$Tingkat\ Akurasi = \frac{1625}{3000}\ X\ 100 = 54.17\%$$

Dapat disimpulkan bahwa dari satu dokumen yang diujikan kepada 30 responden menghasilkan tingkat akurasi sebesar 54.17% terhadap akurasi dokumen ringkasan multibahasa.

5. Kesimpulan

Dapat diimplementasikan system peringkas teks otomatis dokumen tunggal multi bahasa, yaitu bahasa Indonesia dan bahasa Inggris dengan nilai akurasi sebesar 54,17% setelah melalui pengujian pada tiga puluh responden.

Daftar Pustaka

[1] M. Moradi, "CIBS: A biomedical text summarizer using topic-based sentence clustering," *J. Biomed. Inform.*, vol. 88, no. November, pp. 53–61,2018.

[2] M. Mustaqhfiri, Z. Abidin, and R. Kusumawati, "Peringkasan Teks Otomatis Berita Berbahasa Indonesia Menggunakan Metode Maximum Marginal Relevance," *Matics*, no. March 2012, 2012.

[3] N. Grabar and C. Grouin, "A Year of Papers Using Biomedical Texts: Findings from the Section on Natural Language Processing of the IMIA Yearbook," *Yearb. Med. Inform.*, vol. 28, no. 1, pp. 218–222, 2019.

[4] G. Tsatsaronis *et al.*, "An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition," *BMC Bioinformatics*, vol. 16, no. 1, pp. 1–28, 2015.

[5] A. Khan, N. Salim, W. Reafee, A. Sukprasert, and Y. J. Kumar, "A clustered semantic graph approach for multi-document abstractive summarization," *J. Teknol.*, vol. 77, no. 18, pp. 61–72, 2015.