PERINGKASAN KUMPULAN BERITA SECARA OTOMATIS MENGGUNAKAN METODE MAXIMUM MARGINAL RELEVANCE

Filza Setiadi[1*], Esmeralda C Djamal[2], Ridwan Ilyas[3]
[123]Jurusan Informatika, Fakultas Sains dan Informatika, Universitas Jenderal Achmad Yani
Jl.Terusan Jendral Sudirman, Cimahi, Jawa Barat, 40285
[*]Email: filzastd@gmail.com

Abstrak

Berita merupakan suatu informasi yang menceritakan tentang berbagai peristiwa dan kejadian yang hangat, dimana akan berdampak pada penyedia-penyedia berita untuk mengeluarkan berita yang terhangat. Karena hal tersebut banyak ditemukan berita yang sama informasinya dengan judul yang berbeda-beda. Diperlukan metode untuk mengumpulkan informasi ke dalam ringkasan sederhana. Ringkasan berita adalah merupakan teknik untuk mengambil isi yang paling penting dari sumber informasi yang kemudian menyajikannya kembali dalam bentuk yang lebih ringkas. Untuk memecahkan masalah tersebut dapat menggunakan metode maximum marginal relevance. MMR merupakan salah satu metode ekstraksi ringkasan (extractive summary) yang digunakan untuk meringkas dokumen tunggal atau multi dokumen. MMR meringkas dokumen dengan menghitung kesamaan (similarity) antara bagian teks. Pada pengumpulan dan pengambilan data menggunakan teknik web scraping dari portal berita. Tahapan pada teknik web scraping yaitu, mengambil link berita dan disimpan pada sistem, selanjutnya mengambil konten-konten yang diperlukan. Dimana konten yang di ambil adalah konten yang penting, seperti judul berita, isi berita, lokasi berita, serta waktu terbit berita. Pada ekstraksi data terdapat pemecahan kalimat untuk menghilangkan tanda baca dan dipecah lagi menjadi banyak suku kata, pembobotan TF-ISF, consine similaryti, dan pembobotan MMR dalam meringkas suatu kalimat dengan menunjukan frequensi kemunculan kalimat yang yang mendekatkan kata di kaliamat dengan kata kunci.

***Kata kunci**: Berita; Consine Similaryti; Mmr; Penyedia Berita; Tf-Isf.*

1. PENDAHULUAN

Berita merupakan suatu informasi yang menceritakan tentang berbagai peristiwa dan kejadian yang hangat, dimana akan berdampak pada penyedia-penyedia berita untuk mengeluarkan berita yang terhangat. Informasi dalam bentuk teks berita telah menjadi salah satu komoditas yang penting dalam era informasi ini. Sering kali, berita yang sama dituliskan pada berbagai portal. Bahkan dapat disajikan dalam berbagai artikel pada portal yang sama, dengan penambahan sedikit informasi. Hal ini menyebabkan waktu yang diperlukan untuk mendapatkan berita yang sama lebih banyak. Banyak waktu diperlukan untuk menemukan informasi utama dari berita-berita tersebut. Oleh karena itu diperlukan peringkasan kumpulan berita ini agar perolehan informasi dari berita lebih efisien.

Ringkasan berita adalah mengambil isi yang paling penting dari sumber informasi yang kemudian menyajikannya kembali dalam bentuk yang lebih ringkas dengan menjaga konten informasinya. Ringkasan berita sering disebut *multi-document summarization* telah menarik banyak perhatian karena penerapannya dalam aplikasi dunia nyata untuk mendapatkan informasi informal yang diinginkan.(Lukmana, Swanjaya, Kurniawardhani, Arifin, & Purwitasari, n.d.)

Algoritma Maximum Marginal Relevance (MMR) merupakan metode ekstraksi ringkasan yang digunakan untuk meringkas dokumen tunggal maupun multi dokumen (Setiawan & Hartanto, 2016). MMR meringkas dokumen dengan menghitung kesamaan (*similarity*) antara kalimat teks. Pada peringkasan dokumen dengan metode MMR dilakukan proses segmentasi dokumen menjadi kalimat dan dilakukan pengelompokan sesuai dengan jenis kalimat tersebut. MMR digunakan dengan mengkombinasikan matriks *cosine similarity*

untuk merangking kalimat-kalimat sebagai tanggapan pada *query* yang diberikan oleh *user* (Mustaqhfiri, Abidin, & Kusumawati, 2012).

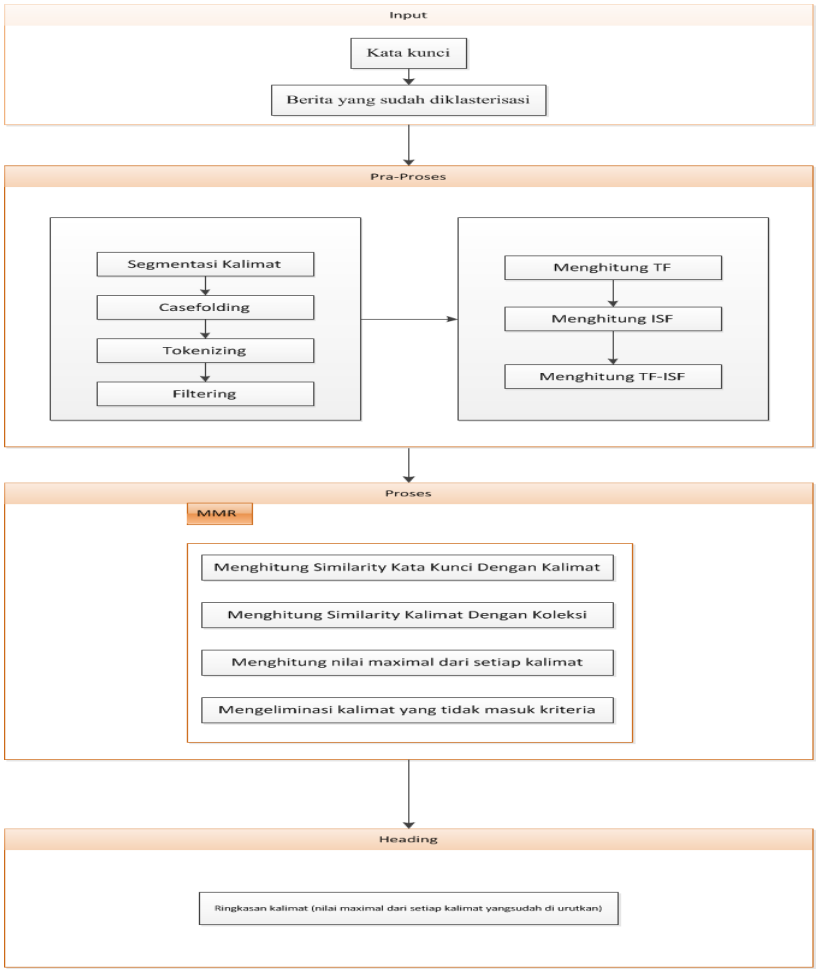
Metode MMR digunakan untuk pemilihan kalimat atau unit teks lain yang memepertimbangkan aspek kerelevanan kalimat dengan *query* dan keterbaruan informasi. Ide dasar dari MMR ini yaitu memberikan penambahan nilai bagi kalimat yang relevan dan memberikan pengurangan nilai redundansi informasi antara kalimat tersebut dengan kalimat lain yang telah terpilih (Goldstein & Carbonell, 1998). Sebuah kalimat dikatakan memiliki *marginal relevance* yang tinggi jika kalimat tersebut relevan terhadap isi dari kalimat dan mempunyai kesamaan bobot *term* maksimum dibandingkan dengan *query* (Marlinda, Rianto, Informatika, Bina, & Informatika, 2013) (Wang, Liu, Sun, Wang, & Li, 2009).

Dari suatu kalimat, perhitungan MMR didasarkan nilai bobot yang diperoleh menggunakan metode Term Frequency Inverse Sentence Frequency (TF-ISF). Penelitian ini telah membuat sistem peringkasan terhadap berita yang sudah diklompokkan sebelumnya. Sistem telah dibangun menggunakan MMR untuk menghitung kedekatan antar kalimat setelah dikonversikan dalam bentuk frekuensi atau bobot melalui proses TF-ISF. Nilai TF-ISF yang digunakan untuk menyatakan bobot hubungan kalimat terhadap *term* (Intan & Defeng, 2006), faktor yang menentukan bobot *term* pada suatu kalimat berdasarkan jumlah kemunculannya dalam kalimat tersebut. Sedangkan, *inverse sentence frequency* merupakan pengurangan dominasi term yang sering muncul diberbagai kalimat (Pratiwi & Bijaksana, 2017).

2. METODE

Algoritma Maximum Marginal Relevance (MMR) merupakan salah satu metode ekstraksi ringkasan (*extractive summary*) yang digunakan untuk meringkas dokumen tunggal atau multi dokumen. MMR meringkas dokumen dengan menghitung kesamaan (*similarity*) antara bagian teks. Pada peringkasan dokumen dengan metode MMR dilakukan proses segmentasi dokumen menjadi kalimat dan dilakukan pengelompokkan sesuai dengan jenis kalimat tersebut. MMR digunakan dengan mengkombinasikan *matrix cosine similarity* untuk merangking kalimat-kalimat sebagai tanggapan pada *query* yang diberikan oleh *user*.

Proses perancangan sistem peringkasan kumpulan berita secara otomatis terdiri dari tahap input, pra proses, proses maximum marginal relevance, dan output. Input pada sistem ini adalah kata kunci dan berita yang sudah terklaster dimana telah di proses dengan engine lain. Selanjutnya tahap pra proses yang terdiri dari tahap ektraksi dokumen dan pembobotan kata. Proses maximum marginal relevance untuk menghasilkan peringkasan kumpulan berita dapat dilihat pada Gambar 1.



Gambar 1. Sistem Optimalisasi Penjadwalan Tour Guuide wisata

3. HASIL DAN PEMBAHASAN

Perancangan sistem dilakukan dengan tiga tahap yaitu, tahap perolehan data, tahap praproses, dan tahap sistem peringkasan kumpulan berita secara otomatis meggunakan metode Maximum Marginal Relevance (MMR). Perancangan sistem peringkasan kumpulan berita secara otomatis.

3.1. Perolehan Data

Portal berita yang di pakai dalam pengambilan data merupakan portal berita kompas dan tempo. Dimana terdapat lima kategori berita yang dipakai. Adapun kategorinya yaitu : olahraga, ekonomi, otomotif, teknologi, dan *entertaintment*. Data diambil menggunakan teknik *scraping*, dimana data diambil dimulai dari tanggal 1 Januari 2018 sampai 30 juni 2018.

3.2. Pra Proses

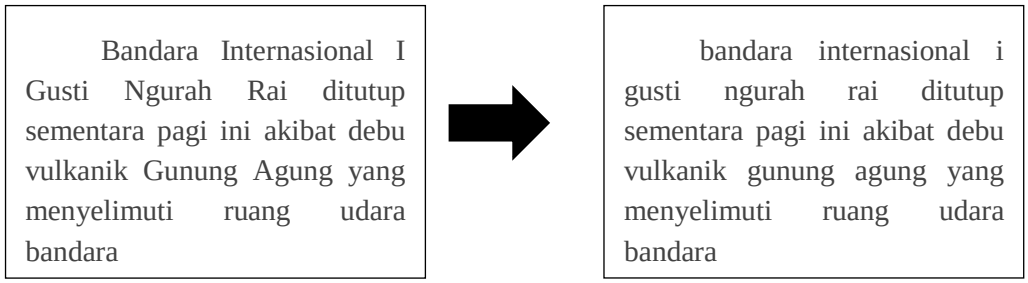
Pra proses menjelaskan tentang ekstraksi dokumen dan pembobotan kata dimana terdiri dari tahap segmentasi, *case folding*, *tokenizing*, *filtering* dan TF-ISF.

3.2.1. Segmentasi

Pada proses segmentasi, dokumen dipecah berdasarkan tanda pemisah kalimat. Setiap dokumen yang telah dipecah akan dimasukkan kedalam list kalimat. Keluaran dari hasil segmentasi berupa kumpulan kalimat yang akan digunakan pada proses berikutnya.

3.2.2. Case Folding

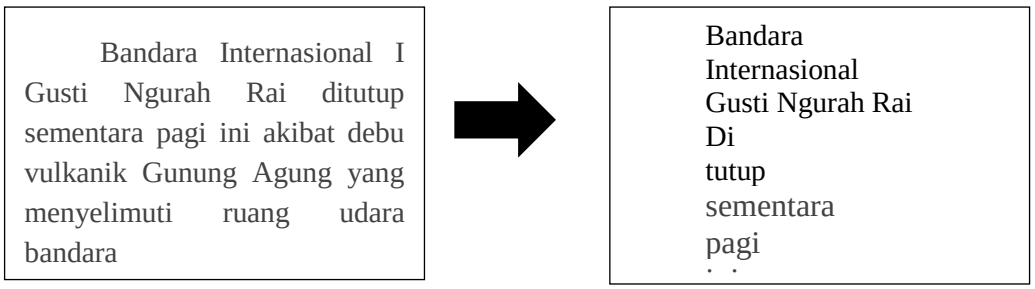
Case folding adalah tahapan proses mengubah semua huruf dalam teks dokumen menjadi huruf kecil, serta menghilangkan tanda baca.



Gambar 2. Case Folding

3.2.3. Tokenizing

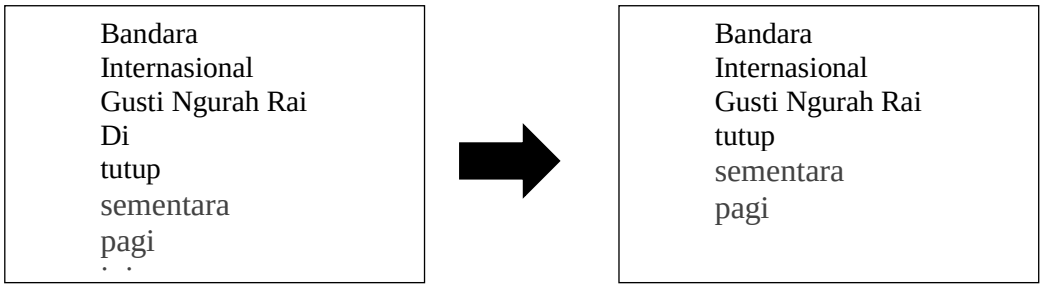
Tokenizing adalah proses pemotongan string input berdasarkan tiap kata yang menyusunnya. Pemecahan kalimat menjadi kata-kata tunggal dilakukan dengan men-*scan* kalimat dengan pemisah (*delimiter*) *white space* (spasi, tab, dan *newline*).



Gambar 3. Tokenizing

3.2.4. Filtering

Setelah dokumen isi berita dipecah per kata, selanjutnya dilakukan proses *filtering* yaitu menghapus kata yang tidak perlu sehingga meringankan proses komputasi. Gambar 5 memperlihatkan proses *filtering*.



Gambar 4. Filtering

3.2.5. Term Frequence Inverse Sentence Frequency

Pembobotan dapat diperoleh berdasarkan jumlah kemunculan suatu *term* dalam sebuah dokumen *term frequency* (*tf*) dan jumlah kemunculan *term* dalam koleksi dokumen *inverse sentence frequency* (*isf*). Bobot suatu istilah semakin besar jika istilah tersebut sering muncul dalam suatu dokumen dan semakin kecil jika istilah tersebut muncul dalam banyak dokumen.

Setelah bobot (W) masing-masing dokumen diketahui, maka dilakukan proses pengurutan (*sorting*) dimana semakin besar nilai W, semakin besar tingkat kesamaan (*similarity)* kalimat tersebut terhadap kata yang dicari, demikian pula sebaliknya.

Terdapat beberapa tahapan untuk mendapatkan nilai bobot diantaranya adalah

a) Menghitung *Term Frequency*

Tabel 1. Tabel Frekuensi Kemunculan Term

Term	Kalimat 1	Kalimat 2	Kalimat 3	Kalimat 4
Term1	0	0	1	0
Term2	0	1	0	0
Term3	1	0	0	2
Term4	1	0	0	1
Term5	0	3	0	2

b) Menghitung Sentence frequency

Tabel 2.Tabel Sentence Frequence

Term	Kalimat Frequency
Term1	1
Term2	1
Term3	2
Term4	4
Term5	5

c) Menghitung invers sentence frequency

Tabel 3. Tabel Inverse Sentence Frequence

Term	Isf
Term1	Log(4/1) = 0,60205
Term2	Log(4/1)= 0,60205
Term3	Log(4/2) = 0,30102
Term4	Log(4/4) = 0
Term5	Log(4/3)= 0,1249

d) Menghitung tf-isf

Tabel 4. Tabel TF-ISF

df	isf		Kalimat 1	Kalimat 2	Kalimat 3	Kalimat 4
1	Log(4/1) 0,60205	=	0	0,60205	0,60205	0
2	Log(4/1) 0,60205	=	0	0,60205	0	0
3	Log(4/2) 0,30102	=	0,30102	0	0	0,60204
4	Log(4/4) = 0		0	0	0	0
5	Log(4/3) 0,1249	=	0	0,3747	0	0,2498
	Sum		0,30102	0.52627	0,60205	0,42592

Maka dapat dilihat dari hasil perhitungan dengan menggunakan metode TF-ISF di atas, dokumen 1 dengan bobot 0,30102, dokumen 2 dengan bobot 1,5788 dokumen 3 dengan bobot 0,60205 dokumen 4 dengan bobot 0,85184

3.3. Maximum Marginal Relevance

MMR utuk peringkasan berita berdasarkan klasterisasi berita memperhatikan setiap proses di antaranya, proses input yaitu klasterisasi yang berisi kumpulan berita yang telah terklaster. Tahap pra-proses, terdiri dari *casefolding*, *tokenizing*, *filtering*, dan pembobotang kata menggunakan TF-ISF. Selanjutnya menghitung kedekatan kata kunci dengan kalimat, lalu menghitung kedekatan antar kalimat, setelah itu menghitung nilai setiap kalimat lalu di masukan dalam koleksi kalimat, dan peroses diulang sampai perhitungan nilai dari kalimat ke n selasai. *Output* dari proses ini adalah kalimat-kalimat yang terpilih untuk menjadi ringkasan.

Setelah mendapatkan hasil bobot query relevance dan matriks bobot *similarity* kalimat, langkah berikutnya adalah melakukan perhitungan pembobotan MMR dengan menggunakan persamaan. Perhitungan MMR dilakukan dengan perhitungan iterasi antara bobot *query relevance* dan bobot *similarity* kalimat. Pada penelitian ini nilai $\lambda = 0.8$. Kalimat yang memiliki nilai MMR paling tinggi dalam setiap iterasi akan terpilih sebagai ringkasan yang pertama dan seterusnya. Iterasi akan berhenti apabila perhitungan MMR menghasilkan nilai 0 (nol) atau minus (-). Hasil MMR untuk setiap iterasi ditunjukkan pada Tabel 3.5.

Tabel 5. Tabel Hasil Iterasi

Ke	K1	K2	K3	K4	K5	K6
1	25,02	-0,03	2,25	4,44	-0,20	2,90
2	-	-0,05	1,73	3,54	-0,04	2,25
3	-	-0,07	1,31	-	-0,05	1,73
4	-	-0,08	0,98	-	-0,06	-
5	-	-0,09	-	-	-0,07	-

Tahap selanjutnya adalah ekstraksi ringkasan dari hasil pembobotan MMR kalimat. Ektraksi ringkasan diperoleh dengan mengambil kalimat yang memiliki nilai MMR paling tinggi untuk setiap itrerasi. Hasil ekstraksi kalimat yang dihasilkan mulai dari nilai MMR tertinggi sampai terendah ditunjukkan pada Tabel 3.6.

Tabel 6. Tabel Hasil Ringkasan

No	Kalimat
K1	Buku kreasi desain produk, distro dan fashion 3D dibuat berdasarkan perkembangan industri-industri di Indonesia, dimana dalam buku ini diajarkan desain-desain produk sederhana, produk yang banyak di produksi di industry skala kecil dan menengah.
K4	Pembahasan diberikan secara lengkap, mulai dari 2D, 3D, hingga operasi rendering
K6	Bagi anda yang sudah cukup ahli, anda dapat langsung belajar mengembangkan desain yang telah ada di dalam CD untuk dibuat menjadi lebih aktif
K3	Banyak lowongan kerja terbuka kerja terbuka bagi anda yang ahli dibidang desain produk
K2	Materi-materi yang diajarkan merupakan bidang keahlian AutoCAD dan 3DS Max yang sangat dibutuhkan oleh industri-industri ditanah air saat ini
K5	Bonus di dalam CD terdapat file-file pendukung dan file latihan

Setelah kalimat diurutkan berdasarkan nilai terbesar, lalu diambil 50% dari hasil iterasi yang telah diurutkan untuk menjadi luaran atau hasil dari ringkasan. Kalimat 1, kalimat 4 dan kalimat 6 adalah hasil yang didapatkan untuk menjadi sebuah ringkasan.

4. KESIMPULAN

Penelitian ini telah menghasilkan sistem yang digunakan sebagai peringkasan kumpulan berita secara otomatis menggunakan metode maximum marginal relevanced. Sistem ini telah diuji melalui dua tahapan proses pengujian yaitu, pengujian sistem dan pengujian akurasi sistem. Pengujian sistem dilakukan dengan menguji fungsionalitas yang disesuaikan dengan rancangan perangkat lunak, sementara itu pengujian akurasi sistem dilakukan untuk menguji akurasi dari sistem yang dibuat dalam ringkasan kalimat. Metode ini dapat digunakan untuk meringkas multi dokumen dengan menggunakan kata kunci (masukan dari user) sebagai query Proses pengujian sistem dilakukan terhadap pengujian *recall, precision, dan f-measure*.

Pengujian tersebut menghasilkan waktu *recall* 69%, *precision* 69% dan menghasilkan *f-measure* sebesar 69%, dengan 10 kali pengujian dan banyak data yai 10 berita.

DAFTAR PUSTAKA

A. Indriani, “Maximum Marginal Relevance Untuk Peringkasan Teks,” pp. 29–34, 2014.

B. S. Pratiwi and M. A. Bijaksana, “Implementasi Word Sense Disambiguation Dengan Metode Maximal Marginal Relevance Pada Peringkasan Teks Implementation of Word Sense Disambiguation Using Maximal,” vol. 4, no. 1, pp. 1152–1157, 2017.

B. Wang, B. Liu, C. Sun, X. Wang, and B. Li, “Multi-email Summarization,” *Work*, pp. 417–424, 2009.

E. B. Setiawan and A. T. Hartanto, “Implementasi Metode Maximum Marginal Relevance (MMR) dan Algoritma Steiner Tree untuk Menentukan Storyline Dokumen Berita,” vol. VIII, no. 1, pp. 23–31, 2016.

I. Lukmana, D. Swanjaya, A. Kurniawardhani, A. Z. Arifin, and D. Purwitasari, “Sentence Clustering Improved Using Topic Words,” pp. 1–8.

J. Goldstein and J. Carbonell, “Summarization:(1) using MMR for diversity-based reranking and (2) evaluating summaries,” *Proc. a Work. held*, no. 1, pp. 181–195, 1998.

L. Marlinda, H. Rianto, M. Informatika, A. Bina, and S. Informatika, “Pembelajaran Bahasa Indonesia Berbasis Web,” pp. 2–4, 2013.

M. Mustaqhfiri, Z. Abidin, and R. Kusumawati, “Peringkasan Teks Otomatis Berita Berbahasa Indonesia Menggunakan Metode Maximum Marginal Relevance,” *Matics*, 2012.

R. Intan and A. Defeng, “Hard : Subject-Based Search Engine Menggunakan Tf-Idf Dan Jaccard ’ S Coefficient,” *J. Tek. Ind.*, vol. 8, no. 1, pp. 61–72, 2006.

S. Irawan, Hermawan, and Samsuryadi, “Studi Awal Peringkasan Dokumen Bahasa Indonesia Menggunakan Metode Latent Semantik Analysis dan Maximum Marginal Relevance,” vol. 2, no. 1, pp. 235–239, 2016.

V. Rentoumi, G. Giannakopoulos, V. Karkaletsis, and G. A. Vouros, “Sentiment Analysis of Figurative Language using a Word Sense Disambiguation Approach,” *Int. Conf. RANLP*, no. September, pp. 370–375, 2009.

W. E. Waliprana and K. Kunci, “Update Summarization Untuk Kumpulan Dokumen

Berbahasa Indonesia," *J. Cybermatika*, pp. 6–10, 2009.