APLIKASI PERINGKASAN DOKUMEN MENGGUNAKANMETODE MAXIMUM MARGINAL RELEVANCE (MMR)

Delvin [1] Desi Arisandi [2] Tri Sutrisno [3]

[1) 2) 3)] Teknik Informatika, FTI, Universitas Tarumanaraga
Jl. Letjen S Parman no 1, Jakarta 11440 Indonesia
[1)] email: Delvin.535170078@stu.untar.ac.id, [2)] email: desia@fti.untar.ac.id, [3)] email: tris@fti.untar.ac.id

ABSTRACT

making a summary application to help readers who do not like to read long and thick news articles which take a relatively long time and can cause readers to be lazy to read the news articles. This summary application is used to summarize news articles, in making the application using ASP.Net. In this summary, the Maximum Marginal Relevance (MMR) method is used. In this study, you can use articles on the website, and do it. Articles are processed in the form of a file (single document) with a txt extension. The summary process goes through the preprocessing stage, which consists of sentence segmentation, case folding, tokenizing, filtering, stemming.

Key words

Article, Maximum Marginal Relevance, Text summary.

1. Pendahuluan

Tujuan pembuatan aplikasi peringkasan dokemen untuk membantu pembaca yang tidak gemar membaca berita atau artikel yang panjang dan tebal yang relatif membutuhkan waktu yang lama dan dapat memyebabkan pembaca malas untuk membaca artikel tersebut. Aplikasi peringkasan ini digunakan untuk meringkas dokumen menggunakan metode *Maximum Marginal Relevance (MMR)*, dalam pembuatan aplikasi tersebut menggunakan ASP.Net. "Peringkasan teks (Automatic Text Summarization) adalah penyortiran beberapa paragraf menjadi bentuk yang lebih singkat menggunakan aplikasi yang dioperasikan dalam komputer (Ajmal dan Haroon, 2016).[1] Algoritma Maximum Marginal Relevance(MMR) merupakan salah satu metode peringkasan yang digunakan untuk meringkas dokumen tunggal. Pada penelitian ini dapat mengunakan artikel yang ada di situs web, dan artikel tersebut akan diringkas menggunakan MMR. Metode Maximum Marginal Relevance(MMR) dipilih agar bisa menyelesaikan permasalahan panjangnya sebuah ceita,karena untuk mendapatkan inti pokok dari informasi tersebut membutuhkan waktu yang lama. Setidak nya pembaca harus membaca halaman demi halaman dari sumber referensi artikel.

2. Peringkasan

Peringkasan teks adalah proses untuk mengambil dan mengekstrak informasi penting dari sebuah teks sehingga menghasilkan teks yang lebih singkat dan menggandung poin-poin penting dari teks sumber. Sebuah sistem peringkasan diberimasukan berupa teks, kemudian melakukan peringkasan, dan menghasilkan keluaran berupa teks yang lebih singkat dari teks aslinya. Pada peringkasan teks terdapat dua pendekatan yaitu, ekstraksi (*shallower approaches*)dan abstraksi (deeper approaches). Pendekatan ekstraksi adalah peringkasan yang memilih suatu paragraf atau kalimat penting dalam menginterprestasikan dokumen kedalam sebuah bentuk sederhana. Sedangkan pendekatan abstraksi menghasilkan ringkasan yang bukan dari kumpulan kalimat penting tetapi menangkap hasildari konsep utama pada teks dan merepresentasikannya menjadi sebuah kalimat baru.[6] Teknik yang digunakan pada penelitian sistem peringkasan teks adalah teknik ekstraksi.:

2.1 Scraping

Scraping ini menggunakan metode xpath, Xpath adalah query language yang bekerja pada dokumen XML. Xpath adalah xml path language. dengan tujuan mencari data dari struktur *file* penunjang halaman. XPath bisa juga digunakan untuk menavigasi struktur dokumen dari dokumen XML yang berbentuk *tree structure* serta memilih nodes dari berbagai parameter. Data yang akan dipakai adalah berita yang di *Scraping* dari suatu website. Untuk data yang digunakan berasal dari website Liputan6.com. URL yang dipakai lansung menujuk ke website yang akan di *Scraping* yang di proses oleh program.

2.2 *Text Preprocessing*

Text Pre-processingmerupakan langkah pra-pemrosesan untuk membuat teks siap untuk diproses. Teks yang ada harus dipisahkan untuk pemrosesan yang lebih mudah, ini dapat dilakukan dalam beberapa level berbeda. Sebuah cerita bisa rusak menjadi paragraf, kalimat, dan pada akhirnya menjadi potongan kata atau disebut token [1]. Dalam aplikasi ini tahapan preprocessing yang digunakan hanya segmentasi kalimat, case folding, tokenization, dan filtering, (menghapus stopword). Stemming tidak digunakan karena akan ada kata kiasan yang tidak sesuai jika stemming dilakukan sehingga mengubah arti kata dalam rangkaian kalimat dalam cerita aslinya. Dalam proses segmentasi kalimat, teks cerita terdiri dari paragraf yang dipecah menjadi beberapa kalimat. Atau bisa jadi walaupun hasil ceritanya hanya 1 paragraf, pemisahannya akan dilakukan dengan memisahkan setiap kalimat berdasarkan tanda baca, seperti titik (.). Kasus Lipat dikonversi semua huruf dokumen menjadi huruf kecil. Hanya huruf 'a' hingga 'z' yang diterima untuk diubah dari huruf besar untuk huruf kecil. Fungsi dari fase ini adalah untuk membuat 2 kata yang sama (walaupun sama) berbeda kapitalisasi/tidak) jika dihitung dalam pengindeksan sehingga menjadi kata yang sama.

Dalam tokenizing adalah fase pemotongan string menjadi token. Proses ini menghilangkan karakter yang melakukan tidak termasuk huruf, seperti tanda baca, kata sambung, angka dan akan dihilangkan. Beberapa kata seperti nama yang diulang tidak akan terputus karena sebuah nama. Proses ini membuatnya mudah dibersihkan kata yang dapat dikenali oleh sistem sehingga dapat menghitung jumlah kata dalam 1 dokumen. Penyaringan adalah untuk menghilangkan kata-kata yang tidak berguna. Stopwords adalah frasa yang kurang deskriptif yang mungkin dibuang dalam pendekatan frase dan mengambil kata-kata penting dari hasil tokenized. Contoh dari Stopword bahasa Indonesia adalah "ini", "yang", "maka", "atas", "ada", "merupakan", "di", dan lain-lain. stopword yang akan digunakan adalah stopword bahasa Indonesia. Pada fase ini, calon daftar kata yang akan dibuang diperlukan agar mereka bisa mendapatkan kata-kata penting yang dibutuhkan untuk perhitungan kata bobot frekuensi sehingga lebih efisien dalam menentukan rangkuman kalimat.

2.3 Term Weighting

Informasi yang diperoleh merupakan kumpulan token dari kata-kata penting. Pengindeksan dilakukan untuk sekelompok kata jadi bahwa mereka dapat mengubah data dalam bentuk kata-kata ini ke dalam bentuk nominal atau numerik sehingga proses perhitungan dapat dilakukan. Metode yang digunakan adalah pembobotan TF-IDF. Istilah Frekuensi (TF) adalah pembobotan kata dengan menghitung total frase dalam suatu dokumen, dari kalimat pertama sampai kalimat terakhir.[2] Inverse Document Frequency (IDF) adalah jumlah kemunculan frase dalam file yang ada koleksi. Namun, karena aplikasi ini hanya memproses satu dokumen, banyak kalimat akan menjadi dihitung secara frekuensi sebagai pengganti jumlah koleksi dokumen, dengan kata lain kalimat merupakan pengganti dokumen.

$$T\,f_{t,d} = \begin{cases} 1 + \log(tf_{t,d}) & if\ tf_{t,d} > 0 \\ 0 & otherwise \end{cases} \quad (1)$$

$$IDF_t = log\ \left(\frac{N}{df_t}\right) \quad (2)$$

$$W_{t,d} = Tf_{t,d} * IDF_t = Tf_{t,d} * log\left(\frac{n}{df_t}\right) \quad (3)$$

2.4 Cosine Similarity

Kesamaan Cosinus dimulai setelah bobot diperoleh, kemudian dihitung ukuran vektornya terlebih dahulu, secara berurutan untuk menghitung kesamaan antara kalimat dan query, serta kesamaan kalimat lain dan kalimat. Semakin besar nilai kemiripan vektor query dengan vektor dokumen (kalimat), maka lebih relevan pertanyaannya dengan kalimat. Kesamaan Cosinus tanpa menormalkan TF-IDF dengan diketahui dengan rumus

$$CosSim(d_j, q) = \frac{\sum_{i=1}^{t}(W_{i,j}.W_{i,q})}{\sqrt{\sum_{i=1}^{t} {W_{i,j}}^2 \cdot \sum_{i=1}^{t} {W_{i,q}}^2}} \quad (4)[1]$$

$$CosSim(d_j, q) = \sum_{i=1}^{t}(W_{i,j} norm\ x\ W_{i,q} norm) \quad (5)$$

2.5 Maximum Marginal Relevance

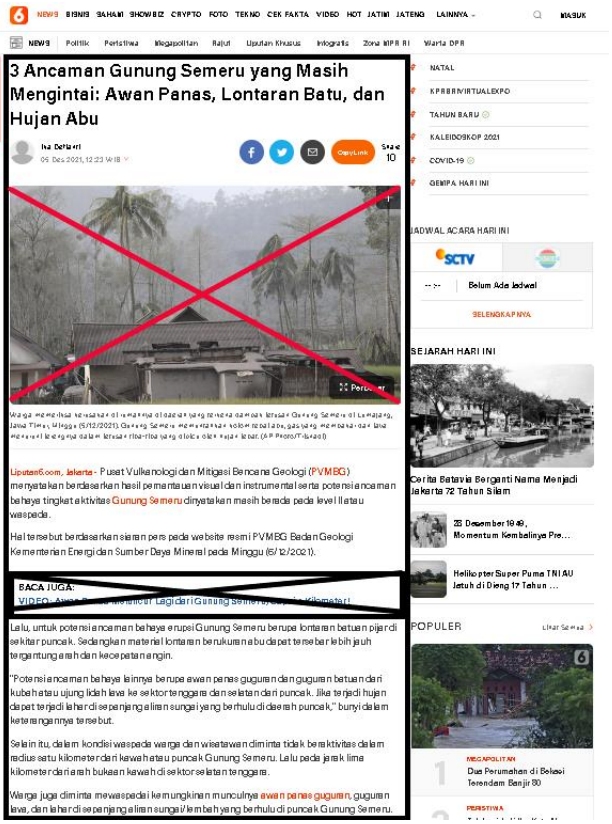
MMR adalah teknik peringkasan yang memiliki tujuan untuk mengambil informasi yang relevan dan tidak mengandung redudansi. MMR meringkas dokumen dengan menghitung kesamaan antara bagian teks dan dengan tujuan mendapatkan skor kalimat berdasarkan kesamaan (similarity) dengan query yang diberikan dan dapat mengurangi redudansi pada hasil ringkasan yang di dapat.Metode Maximum Marginal Relevance (MMR) sering digunakan untuk peringkasan teks karena metode MMR sederhana dan efisien (Xie & Liu, 2008)[3]. Jika kesamaan antara satu kalimat dengan kalimat yang lain tinggi, maka terdapat kemungkinan terjadi redudansi. Metode MMR dapat mengurangi redundasi dengan rumus pada Persamaan

$$MMR = argmax[\lambda * Sim1(Si, Q) - (1 - \lambda) * \max Sim2(Si, S')] \quad (6)$$

3. Hasil Percobaan

Data yang akan dipakai adalah berita yang di *Scraping* dari suatu website. Untuk data yang digunakan berasal dari website Liputan6.com. URL yang dipakai lansung menujuk ke website yang akan di *Scraping* yang di proses oleh program

[1] Nirmala Fa'izah Saraswati , Indriati , Rizal Setya Perdana., op.cit., h.4.



Gambar 1 Website berita di liputan6

Tabel 1 Perbandingan berita asli dan peringkasan

no	Berita
1	Pusat Vulkanologi dan Mitigasi Bencana Geologi (PVMBG) menyatakan berdasarkan hasil pemantauan visual dan instrumental serta potensi ancaman bahaya tingkat aktivitas Gunung Semeru dinyatakan masih berada pada level II atau waspada. Hal tersebut berdasarkan siaran pers pada website resmi PVMBG Badan Geologi Kementerian Energi dan Sumber Daya Mineral pada Minggu (5/12/2021). Lalu, untuk potensi ancaman bahaya erupsi Gunung Semeru berupa lontaran batuan pijar di sekitar puncak. Sedangkan material lontaran berukuran abu dapat tersebar lebih
2	hal tersebut berdasarkan siaran pers pada website resmi pvmbg badan geologi kementerian energi dan sumber daya mineral pada minggu (5/12/2021) pusat vulkanologi dan mitigasi bencana geologi (pvmbg) menyatakan berdasarkan hasil pemantauan visual dan instrumental serta potensi ancaman bahaya tingkat aktivitas gunung semeru dinyatakan masih berada pada level ii atau waspada

Dari data survey yang disebarkan menggunakan *Question and Answering* yang berupa kuesioner. Didapatkan total responden sebanyak 27 orang. Responden diminta untuk menjawab 5 pertanyaan dari tiap ringkasan sehinga total yang harus di jawab sebanyak 15 pertanyaan.

Tabel 2 Akurasi MMR

Ya Ringkasan 1	**Ya Ringkasan 2**	**Ya Ringkasan 3**
101	82	93
Avg ringkasan 1	**Avg ringkasan 2**	**Avg ringkasan 3**
74.815%	60.7408%	68.889%
AVG Total MMR		
68.148%		

Keusioner *google* fomulir. Pengujian ini mendapatkan 27 responden yang menjawab pertanyaan-pertanyaan mengenai hasil dari ringkasan MMR. Tiap ringkasan memiliki 5 buat pertanyaan untuk responden. Untuk hasil dari akurasi rigkasannya itu sendiri didapatkan persentase yaitu 68.15%. ini merupakan nilai yang cukup baik untuk hasil pengujian dari ringkasan tersebut

4. Kesimpulan

1. Pada pengujian peringkasan dari ketiga ringkasan tersebut dapat dilihat bahwa nilai akurasi dari MMR adalah 68,15%.
2. Pendapatan nilai akurasi tersebut dari *Question and Answering (QnA).*
3. Batasan dari peringkasan ini ditentukan dari perolehan nilai MMR ketika nilai dari MMR maksimum dari tiap kalimat tersebut lebih besar dari 0 maka akan dijadikan ringkasan..

REFERENSI

[1] Saraswati, Nirmala Fa'izah; Indriati; dan Perdana, Rizal Setya; "Peringkasan Teks Otomatis Menggunakan Metode Maximum Marginal Relevance Pada Hasil Pencarian Sistem Temu Kembali Informasi Untuk Artikel Berbahasa Indonesia". Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. Vol. II, Nomor 11. November 2018.

[2] Wahyudi, Dwi; Susyanto, Teguh; dan Nurgroho, Didik. " Implementasi dan Analisis Algoritma Stemming Nazief & Adriani dan Porter pada Dokumen Berbahasa Indonesia". Jurnal Ilmiah Sinus. Vol. XV, Nomor 2.

[3] Savayona, Eva "PERINGKASAN DOKUMEN BAHASA INDONESIA PADA CERPEN MENGGUNAKAN METODE MAXIMUM MARGINAL RELEVANCE" Sumber : Dok baru 2019-08-28 23.05.59 (unsri.ac.id).

[4] Setiawan, Eko Budi dan Hartanto, Aji Teja. "Implementasi Metode Maximum Marginal Relevance (MMR) dan Algoritma Steiner Tree untuk Menentukan

Storyline Dokumen Berita". ULTIMATICS. Vol. VIII, Nomor 1. Juni 2016.

Delvin, mahasiswa S1, program studi Teknik Informatika, Fakultas Teknologi Informasi Universitas Tarumanagara.