Jurnal Linguistik Komputasional
Indonesian Association for Computational Linguistics
JLK
Vol. 1, No. 1
Maret 2018

Peringkasan Multidokumen Otomatis dengan Menggunakan *Log-Likelihood Ratio* (LLR) dan *Maximal Marginal Relevance* (MMR) untuk Artikel Bahasa Indonesia

Ikhwan Nizwar Akhmad[#1], Anto Satriyo Nugroho[*2], Bambang Harjito[#3]

[#]*Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sebelas Maret*
Jl. Ir. Sutami No. 36A, Surakarta, Indonesia

[1]ikhwan.nizwar@student.uns.ac.id

[3]bambang_harjito@staff.uns.ac.id

[*]*Pusat Teknologi Informasi & Komunikasi, Badan Pengkajian dan Penerapan Teknologi*
Gedung Teknologi 3, Lt. 3, Puspitek Serpong, Indonesia

[2]asnugroho@ieee.org

Abstract* - *Increasing number of information available on the Internet, along with its benefit, also comes with various problems. Modern search engines are smart enough to bring the most relevant information, but the immense number of information provided often brings more confusion than clarity. This condition is known as information overload. Automatic multidocument summarization is a way to overcome this particular problem. Nevertheless, despite of being heavily studied more than 20 years, its implementations for Indonesian language are limited. In this paper, we reported our experimental results on multidocument summarization in Indonesian language. Articles about infectious disease is one of the ideal case study for multidocument summarization for Indonesian language. Information about infectious disease are essential for general public therefore many information about it is available on the Internet. This condition could trigger information overload when someone do an internet search in this topic.

In this research, we try to implement multidocument summarization technique for articles with infectious disease topic in Bahasa Indonesia utilizing Log Likelihood Ratio (LLR) to obtain topic signatures and Maximal Marginal Relevance (MMR) to generate relevant summary with minimal information redundancy. Our summarization method generated a summary with 0.4 F-measure using ROUGE-S9 evalution. Also, we found that topic signature (with its accuracy) takes an important role on generating good summaries.

Keywords* - *multidocument summarization, topic signatures generation

Abstrak - Peningkatan jumlah informasi yang tersedia di internet disamping memberikan manfaat, juga memunculkan masalah tersendiri. Mesin pencarian modern sudah cukup baik untuk mendapatkan informasi tertentu. Namun jumlah informasi yang banyak terkadang menyebabkan pencari informasi kesulitan mendapatkan intisari dari informasi yang dicari. Kondisi ini dikenal sebagai *information overload*. Peringkasan multidokumen otomatis adalah salah satu solusi untuk masalah ini. Meskipun metode peringkasan multidokumen otomatis sudah dikembangkan sejak 20 tahun lalu, penerapannya dalam Bahasa Indonesia masih terbatas. Dalam tulisan ini, kami melaporkan hasil penelitian yang dilakukan pada peringkasan multidokumen berbahasa Indonesia. Artikel dengan topik penyakit menular merupakan salah satu studi kasus yang menarik untuk peringkasan multidokumen Bahasa Indonesia. Informasi mengenai penyakit menular dibutuhkan oleh masyarakat sehingga tersedia banyak informasi mengenai topik ini di internet. Kondisi ini menyebabkan kemungkinan information overload untuk pencarian dalam topik ini.

Dalam penelitian ini, diterapkan peringkasan multidokumen otomatis dengan menggunakan *Log-Likelihood Ratio* (LLR) untuk mendapatkan *topic signature*, dan *Maximal Marginal Relevance* pada artikel dengan topik penyakit menular untuk mendapatkan ringkasan dengan sedikit perulangan informasi. Penelitian ini menghasilkan ringkasan dengan nilai akurasi sebesar 0,4 (dengan menggunakan ROUGE-S9). Selain itu, dalam penelitian ini didapatkan bahwa *topic signature* (beserta akurasinya)

Submitted 24-01-2018; Revised 08-03-2018; Accepted 15-03-2018

memegang peran penting dalam proses peringkasan dokumen otomatis.

Kata Kunci - peringkasan multidokumen otomatis, *topic signature* generation

I. PENDAHULUAN

Penggunaan internet dalam kehidupan sehari-hari di Indonesia selalu meningkat setiap saat. Terdapat 72,7 juta pengguna aktif internet di Indonesia pada tahun 2015 [1]. Pada tahun 2016, jumlah pengguna aktif internet di Indonesia mencapai 88,1 juta pengguna [2]. Sampai dengan awal tahun 2017, pengguna aktif internet di Indonesia telah mencapai 132,7 juta [3]. Seiring dengan era Web 2.0 dimana pengguna menjadi salah satu penyedia konten dan informasi di internet [4], peningkatan jumlah pengguna akan menyebabkan peningkatan jumlah konten yang tersedia di internet.

Dampak yang disebabkan oleh semakin banyaknya konten yang tersedia di internet adalah information overload. Information overload adalah istilah yang dipakai untuk mewakili suatu kondisi dimana efisiensi seseorang dalam melakukan suatu pekerjaan berkurang akibat jumlah informasi relevan dan berpotensi bermanfaat terlalu banyak. Kondisi ini dapat menyebabkan stres, dan kurang fokus [5].

Salah satu solusi untuk mengatasi information overload adalah dengan meringkas beberapa informasi sekaligus. Meskipun proses peringkasan dapat dilakukan secara manual oleh manusia, proses ini sangat menguras tenaga, waktu, dan biaya. Oleh karena itu, peringkasan otomatis perlu dilakukan.

Penelitian mengenai metode peringkasan otomatis telah dilakukan sejak 20 tahun yang lalu, baik untuk peringkasan dokumen tunggal ataupun peringkasan multidokumen. Kebanyakan peringkasan dilakukan terhadap dokumen Bahasa Inggris, sementara beberapa dilakukan untuk bahasa selain Bahasa Inggris. Meskipun penelitian topik ini telah begitu banyak dilakukan, penerapannya terhadap Bahasa Indonesia masih terbatas. Penerapan metode peringkasan multidokumen otomatis yang sudah dikembangkan untuk Bahasa Indonesia adalah dengan metode *sentence scoring* [6]–[8].

Topik mengenai penyakit menular merupakan salah satu topik menarik yang dapat digunakan topik peringkasan otomatis karena merupakan salah satu informasi yang dibutuhkan oleh masyarakat dan internet seringkali menjadi rujukan pertama untuk mendapatkannya. Kebutuhan ini menyebabkan banyaknya jumlah informasi yang tersedia. Metode peringkasan otomatis diperlukan agar informasi mengenai topik ini mudah diserap oleh manyarakat.

Dalam penelitian ini, akan diterapkan metode peringkasan otomatis multidokumen dengan memanfaatkan Log Likelihood Ratio (LLR) dan *Maximal Marginal Relevance* (MMR) untuk menghasilkan ringkasan yang sesuai topik dengan perulangan informasi yang minimal.

II. PERINGKASAN DOKUMEN OTOMATIS

Peringkasan dokumen otomatis telah menjadi salah satu bidang yang diteliti oleh komunitas Natural Language Processing (NLP) lebih dari setengah abad yang lalu. Terdapat beberapa jenis peringkasan dokumen otomatis: berdasarkan jumlah dokumen yang diringkas (tunggal vs multidokumen), metode pembentukan kalimat ringkasan (ekstraktif vs abstraktif), dan ada tidaknya kata pencarian.

A. Peringkasan Dokumen Tunggal vs Multidokumen

Metode peringkasan dokumen dibedakan berdasarkan jumlah dokumen yang diringkas. Penelitian awal pada metode peringkasan dokumen lebih banyak dilakukan untuk dokumen tunggal. Pada awal perkembangannya, metode yang paling sering dipakai adalah dengan memanfaatkan fitur dokumen untuk menghitung significant value sebagai acuan pemilihan kalimat ringkasan. Beberapa fitur yang digunakan diantaranya word frequency, posisi kalimat, cue word, dan kategori kerangka dokumen [9]–[11]. Metode yang lebih maju memanfaatkan machine learning untuk memilih kalimat, diantaranya dengan menggunakan naive bayes classifier dan decision tree [12], [13].

Metode peringkasan multidokumen memiliki perbedaan dengan peringkasan dokumen tunggal dikarenakan adanya informasi yang sama, saling melengkapi, atau berlawanan yang tersebar pada dokumen-dokumen yang akan diringkas. Oleh karena itu, peringkasan otomatis dokumen memiliki fokus tidak hanya menghindari perulangan informasi dalam ringkasan namun juga mengenali informasi baru dan memastikan hasil akhir ringkasan lengkap dan koheren [14].

McKeown dan Radev [15] mengusulkan metode peringkasan multidokumen dengan konsep clustering kalimat dengan memanfaatkan similaritas antar kalimat dari beberapa dokumen yang akan diringkas. Carbonel dan Goldstein [16] mengusulkan metode dengan memanfaatkan *Maximal Marginal Relevance* (MMR) untuk menilai kebaruan informasi pada suatu kalimat. Meskipun peringkasan otomatis dokumen telah banyak dikembangkan, implementasi metode dalam Bahasa Indonesia masih belum banyak diterapkan.

Di antara berbagai penelitian mengenai peringkasan multidokumen pada bahasa Indonesia, banyak yang memakai metode *sentence scoring*. Metode ini memanfaatkan fitur kalimat seperti *word frequency (wf), term frequency-inverse document frequency (td-idf)*, posisi kalimat, kemiripan kalimat terhadap judul, dan kemiripan kalimat terhadap *trending issue* [6]–[8]. Dalam studi ini, peringkasan multidokumen dilakukan dengan memanfaatkan topic signature. Akan tetapi pemilihan

Korespondensi : Ikhwan Nizwar Akhmad

dilakukan dengan memakai nilai Log Likehood Ratio (LLR) tertinggi, karena pendekatan Gupta et.al. [18] yang berbasis cut-off 10 menghasilkan topic signature yang terlalu banyak sehingga tidak dapat menghasilkan ringkasan yang bagus.

B. Metode Peringkasan Ekstraktif vs Abstraktif

Pada metode peringkasan ekstraktif, kalimat dalam hasil ringkasan adalah sebagian kalimat dari dokumen yang akan diringkas. Metode ekstraktif dilakukan dengan memilih kalimat dari dokumen asal untuk dimasukkan ke dalam ringkasan. Sedangkan pada metode peringkasan abstraktif, kalimat ringkasan dibuat dari nol berdasarkan informasi yang didapat dari dokumen yang diringkas (dengan memanfaatkan metode pembentukan kalimat).

Meskipun pada awalnya metode ekstraktif dianggap kurang baik untuk peringkasan multidokumen, anggapan ini berubah ketika sistem peringkasan MEAD yang dikembangkan Radev et.al. [17] memberikan performa yang baik dengan metode ekstraktif. Metode abstraktif akan memberikan ringkasan dalam susunan yang lebih baik, namun sangat sulit diimplementasikan tanpa tersedianya metode pembentukan kalimat yang baik [14].

C. Kata Pencarian dalam Peringkasan Dokumen Otomatis

Artikel (selain artikel tanya jawab) biasanya ditulis untuk memberikan sejumlah informasi kepada pembaca tanpa memperhatikan informasi mana yang diperlukan oleh pembaca. Keberadaan kata pencarian mempengaruhi metode peringkasan tergantung pada tujuan suatu sistem peringkasan dibuat. Dengan tidak adanya kata pencarian, ringkasan akan merefleksikan informasi yang ingin disampaikan penulis dari dokumen yang akan diringkas [14]. Kata pencarian menjadi penting dalam metode peringkasan ketika ringkasan perlu disajikan sebagai jawaban terhadap suatu pertanyaan oleh pencari informasi.

D. Topic Signature dalam Peringkasan Dokumen Otomatis

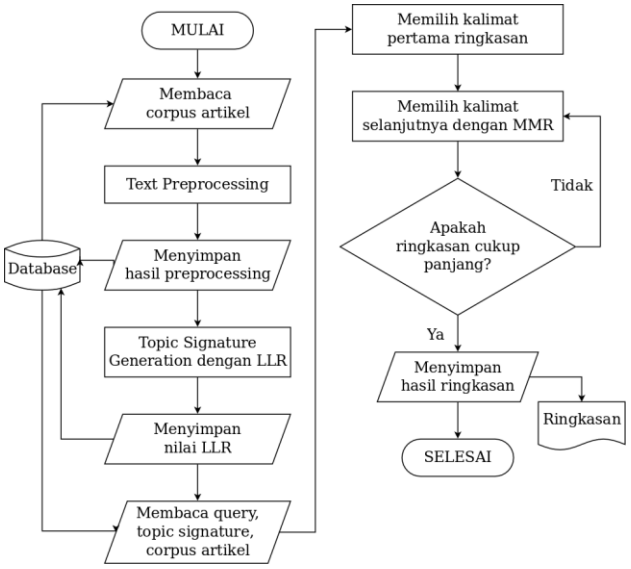
Kebutuhan seseorang dalam pencarian informasi biasanya berpusat pada suatu topik tertentu. Dokumen di internet biasanya memiliki topik tertentu dan setiap topik memiliki informasi penting yang berbeda-beda. Pada dokumen dengan topik mengenai kejadian sejarah atau bencana alam, waktu dan tanggal merupakan informasi yang sangat penting sedangkan pada dokumen dengan topik mengenai ulasan restoran, informasi yang sama menjadi kurang penting.

Topic Signature adalah kata-kata penting dan sangat relevan terhadap suatu topik tertentu [18]. Sebagai contoh, *topic signature* untuk topik "penyakit menular" antara lain "tular", dan "virus". *Topic signature* dapat dijadikan acuan bahwa sebuah dokumen teks relevan terhadap suatu topik, dan suatu kalimat mengandung informasi penting terhadap suatu topik tertentu.

Topic signature dapat dipilih secara manual untuk suatu topik, namun hal ini tidak mungkin dilakukan mengingat jumlah topik akan terus bertambah. Untuk mengatasi kondisi ini, Aone et. al. memanfaatkan tf-idf (term frequency and inverse document frequency) untuk memilih signature words yang kemudian digunakan sebagai salah satu fitur dalam peringkasan otomatis [12]. Lin dan Hovy [19] menggunakan log likelihood ratio (LLR) untuk memilih *topic signature* secara otomatis untuk sistem peringkasan otomatis. Pada penelitian lanjutan, ditemukan bahwa pemilihan *topic signature* dengan memanfaatkan LLR memberikan hasil ringkasan yang lebih baik dibandingkan dengan menggunakan tf-idf terutama untuk peringkasan dokumen dengan topik dan kata pencarian [18].

III. METODE YANG DIUSULKAN

Metode peringkasan yang diusulkan terbagi menjadi dua bagian; *topic signature generation* untuk topik penyakit menular dengan memanfaatkan *Log-Likelihood Rasio* (LLR), dan pembentukan ringkasan multidokumen dengan menggunakan *Maximal Marginal Relevance* (MMR). Metode peringkasan yang akan dilakukan dijelaskan pada Gambar 1.



Gambar 1 Metode Peringkasan

A. Database Korpus Artikel

Pada penelitian ini, akan digunakan korpus artikel yang berisi 842 artikel dengan topik penyakit. Artikel didapatkan dari internet (doktersehat.com dan alodokter.com, diakses antara 27 Februari s.d. 03 Maret 2017) dengan sedikit modifikasi format artikel. Dalam korpus tersebut terdapat 119 (14%) artikel dengan topik penyakit menular, sebagai topik yang akan digunakan pada penelitian ini.

B. Text Preprocessing

Sebelum proses peringkasan, dilakukan text preprocessing terhadap setiap artikel dalam korpus. Text

preprocessing dilakukan untuk mempersiapkan data untuk proses peringkasan. Pada metode yang diusulkan, text preprocessing terbagi menjadi tiga bagian: tokenization, stop word removal, dan stemming.

Tokenization: Artikel-artikel dalam korpus dipisah-pisahkan berdasarkan kalimat menjadi kumpulan kalimat. Kalimat kemudian dipisah-pisahkan kembali menjadi kumpulan kata.

Stop word removal: Stop word adalah kata yang dianggap tidak memiliki nilai informasi. Biasanya merupakan kata yang sangat sering muncul dalam suatu bahasa (contoh: "yang", "karena", dan "dan"). Stop word removal dilakukan dengan membuang kalimat yang terdapat pada daftar stop word bahasa indonesia. Daftar stop word bahasa indonesia didapat dari [20].

Stemming: pengembalian suatu kata ke dalam bentuk dasarnya (contoh: "menyapu menjadi sapu"). Proses stemming dalam penelitian ini akan mengunakan metode confix stripping oleh Mirna et.al [21].

Hasil dari keseluruhan proses text preprocessing disimpan ke database untuk proses selanjutnya.

C. *Topic Signature Generation dengan LLR*

Lin dan Hovy [19] mengusulkan pemilihan *topic signature* secara otomatis dengan menggunakan *Log-Likelihood Ratio* (LLR) suatu kata dalam kumpulan dokumen (korpus). Beberapa kata dengan LLR tertinggi digunakan sebagai *topic signature* suatu topik yang termasuk dalam korpus. Untuk melakukan penghitungan LLR, terlebih dahulu dokumen dalam korpus dikelompokkan menjadi dua kelompok: dokumen yang relevan terhadap topik, dan dokumen yang tidak relevan terhadap topik.

Misalkan pada Tabel 1, adalah daftar kata (unik) dari korpus. Untuk dengan adalah jumlah kata (unik) dalam korpus, merupakan suatu kata dalam dengan indeks . adalah himpunan artikel relevan dengan topik, sedangkan adalah himpunan artikel yang tidak relevan.

Tabel 1 LLR Contingency Table

	R	$\bar{R}$
$C(t_i)$	$k1_i$	$k2_i$
Σt	$n1$	$n2$

Pada Tabel 1, menyatakan jumlah kemunculan suatu kata dalam korpus. adalah jumlah kemunculan kata pada kelompok dokumen relevan. adalah jumlah kemunculan dalam kelompok dokumen tidak relevan.

$$\Sigma t = \sum_{i=1}^{m} C(t_i) \quad (1)$$

merupakan jumlah kata dalam korpus, sesuai dengan Persamaan (1). adalah jumlah kata dalam kumpulan dokumen relevan, dan adalah jumlah kata dalam kumpulan dokumen tidak relevan.

$$L(p,k,n) = p^k(1-p)^{n-k} \quad (2)$$

$$p1_i = \frac{k1_i}{n1}; p2_i = \frac{k2_i}{n2}; p_i = \frac{k1_i + k2_i}{n1+n2} \quad (3)$$

Persamaan (2) digunakan untuk menghitung likelihood suatu kata. Dengan menggunakan persamaan (3) untuk menghitung probabilitas tiap kata berdasarkan nilai yang didapat dari Tabel 1, LLR untuk suatu kata dapat dihitung dengan menggunakan persamaan (4).

$$\begin{aligned} LLR(t_i) &= -2log\lambda \\ &= 2[logL(p1_i, k1_i, n1) \\ &\quad + logL(p2_i, k2_i, n2) \\ &\quad - logL(p_i, k1_i, n1) \\ &\quad - logL(p_i, k2_i, n2)] \end{aligned} \quad (4)$$

Nilai LLR dihitung untuk setiap kata yang terdapat dalam korpus. Selanjutnya kata diurutkan berdasarkan nilai LLR-nya. *Topic signature* didapat dengan memilih beberapa kata yang memiliki nilai LLR tertinggi.

Metode ini kemudian dikembangkan oleh Gupta et.al. [18] dengan menambahkan nilai cut-off untuk menentukan kata yang termasuk sebagai *topic signature* (sehingga tidak diperlukan pengurutan nilai LLR untuk mendapatkan *topic signature*).

D. *Peringkasan Otomatis Multidokumen dengan MMR*

Redundansi informasi merupakan salah permasalahan yang muncul dalam peringkasan multidokumen. Carbonel dan Goldstein [16] menggunakan *Maximal Marginal Relevance* (MMR) dalam pemilihan kalimat ringkasan untuk mengurangi redundansi informasi. MMR dihitung dengan mengukur relevansi kalimat kandidat ringkasan dengan kata pencarian, kemudian dan membandingkan kalimat tersebut dengan kalimat yang yang sudah terpilih untuk masuk dalam ringkasan. Persamaan (5) digunakan untuk melakukan proses pemilihan kalimat menggunakan MMR.

$$\begin{aligned} MMR \equiv \arg \max_{D_i \in R \setminus S} [\lambda(Sim_1(D_i, Q)) - (1 \\ - \lambda) \max_{D_j \in S} Sim_2(D_i, D_j)] \end{aligned} \quad (5)$$

adalah similaritas antara kalimat kandidat ringkasan (berasal dari korpus) dengan kata pencarian. adalah similaritas kalimat kandidat ringkasan dengan kalimat yang sudah terpilih dalam ringkasan. digunakan untuk mengatur preferensi antara relevansi kalimat kandidat dengan kata pencarian dan kesamaan informasi dalam kalimat kandidat dengan kalimat yang sudah terpilih dalam ringkasan. akan memberikan kalimat yang sangat relevan dengan kata pencarian namun memiliki kesamaan informasi dengan kalimat yang dipilih sebelumnya. akan memilih kalimat dengan kesamaan informasi yang rendah namun tidak relevan dengan kata pencarian.

1) *Pemilihan Kalimat Pertama:* Proses peringkasan otomatis multidokumen dengan memanfaatkan MMR tidak dapat digunakan pada pemilihan kalimat pertama. Sehingga metode yang diusulkan akan memilih kalimat pertama kandidat ringkasan tanpa menggunakan MMR. Pemilihan kalimat pertama dilakukan dengan menilai kalimat dalam korpus berdasarkan *topic signature* dan kata pencarian. Kalimat yang didalamnya terdapat kata yang termasuk dalam *topic signature* atau kata pencarian paling banyak akan dipilih sebagai kalimat pertama ringkasan.

2) *Cosine Similarity:* Sesuai dengan Persamaan (5), MMR membutuhkan metode untuk mengukur similaritas antara dua kalimat. Cosine similarity merupakan salah satu metode yang bisa diterapkan untuk mengukur similaritas teks. Teks terlebih dahulu diubah menjadi vektor berdasarkan pada jumlah kata yang terdapat pada masing-masing kalimat. Persamaan (6) digunakan untuk mengukur similaritas antara dua vektor dan .

$$Sim(A,B) = \frac{A \cdot B}{\| A \| \| B \|} \quad (6)$$

E. Evaluasi Metode Peringkasan Dokumen Otomatis

Evaluasi peringkasan dokumen otomatis dapat dilakukan dengan pengamatan manusia dengan berbagai ukuran kualitas seperti koherensi, kejelasan, struktur bahasa, keterbacaan, dan kandungan isinya [12]. Namun proses evaluasi secara manual dengan pengamatan manusia merupakan pekerjaan yang sangat menguras waktu, tenaga, dan biaya sehingga tidak mungkin dilakukan dalam penelitian yang membutuhkan evaluasi metode yang perlu diulang setiap kali terdapat perbaikan.

Lin [22] mengusulkan metode peringkasan otomatis dengan memanfaatkan statistik kata dan dan frasa dalam ringkasan. Metode ini disebut dengan Recall-Oriented Understudy for Gisting Evaluation (ROUGE). ROUGE menggunakan ringkasan manual yang dibuat manusia sebagai referensi untuk menilai akurasi suatu ringkasan yang dihasilkan dengan menggunakan metode otomatis. ROUGE mendapatkan akurasi suatu ringkasan dengan memanfaatkan unit yang muncul bersamaan pada dua ringkasan yang dibandingkan. Unit yang digunakan bisa berupa kata, bigram (dan n-gram lainnya), barisan kata, dan pasangan kata.

Evaluasi yang dilakukan untuk menguji metode ROUGE menunjukkan bahwa skip bigram (bigram yang memperbolehkan jarak antara dua kata) adalah salah satu unit yang baik untuk menguji metode peringkasan otomatis multi-dokumen.

$$R_{skip2} = \frac{SKIP(RM,RO)}{Count_{skip2}(RM)} \quad (7)$$

$$R_{skip2} = \frac{SKIP(RM,RO)}{Count_{skip2}(RO)} \quad (8)$$

$$F_{skip2} = \frac{(1+\beta^2)R_{skip2}P_{skip2}}{R_{skip2} + \beta^2 P_{skip2}} \quad (9)$$

Pada Persamaan (7) dan (8) adalah ringkasan manual yang dibuat oleh manusia, dan adalah ringkasan dari metode peringkasan otomatis. Persamaan (9) adalah fungsi yang digunakan untuk mengukur akurasi ringkasan otomatis terhadap ringkasan yang dibuat manusia.

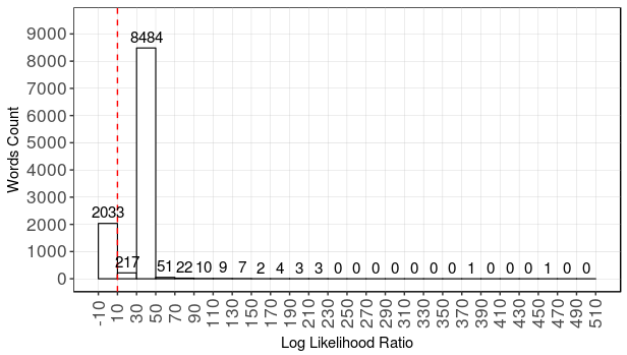
Evaluasi dilakukan dengan menghitung akurasi hasil ringkasan dengan menggunakan ROUGE. Evaluasi berulang dengan mengubah parameter jumlah *topic signature*, bobot dalam MMR, dan minimum word dalam pemilihan kalimat.

IV. Hasil dan Pembahasan

A. Topic Signature Generation

Topic signature untuk topik penyakit menular yang digunakan untuk peringkasan otomatis didapat dengan menghitung nilai LLR setiap kata yang terdapat dalam korpus. *Topic signature* kemudian dipilih dengan memilih sejumlah kata dengan LLR tertinggi [19] atau dengan menggunakan nilai cutting point [18].

Dari seluruh artikel dalam korpus, setelah text preprocessing didapat 10.848 kata unik. Nilai LLR untuk setiap kata dihitung dengan menggunakan topik penyakit menular sebagai kelompok dokumen relevan dan sisanya sebagai kelompok dokumen tidak relevan. Nilai LLR masing-masing kata bervariasi dari -8,87 s.d. 516,95. Gambar 2 merupakan histogram yang menggambarkan nilai LLR dan jumlah kata dengan nilai LLR dalam interval.



Gambar 2 Histogram LLR

Berdasarkan pada Gambar 2, jika cutting point dipakai untuk penentuan *topic signature*, akan didapat 8.815 kata yang digunakan *topic signature*. Oleh karena jumlah topic signature terpilih yang terlalu besar, penentuan *topic signature* dilakukan dengan menentukan jumlah *topic signature* dan memilih sejumlah kata dengan LLR tertinggi. Jumlah kata yang digunakan sebagai *topic*

signature ditentukan dengan melakukan pengujian terhadap hasil ringkasan.

Tabel 2 memuat 10 kata dengan nilai LLR tertinggi berdasarkan korpus yang digunakan dalam penelitian ini.

Tabel 2 Kata dengan LLR tertinggi

No.	Kata	LLR
1	tular	516,95
2	virus	450,68
3	infeksi	373,68
4	kanker	227,78
5	cacing	223,48
6	flu	217,41
7	nyamuk	205,07
8	hewan	192,07
9	tinea	191,74
10	kutu	189,75

B. Pemilihan Kalimat Pertama Ringkasan

Kalimat pertama dalam ringkasan dipilih dengan melakukan penilaian kalimat dalam korpus berdasarkan kata pencarian dan *topic signature*. Metode ini digunakan dengan harapan mendapatkan kalimat pertama ringkasan yang berasal dari dokumen dengan topik penyakit menular dan sesuai dengan kata pencarian. Meski demikian, metode ini sangat bergantung pada akurasi *topic signature* terhadap suatu topik. Pengaruh jumlah *topic signature* yang digunakan terhadap kalimat pertama yang terpilih dijelaskan pada Tabel 3.

Tabel 3 Pemilihan kalimat pertama pada ringkasan

Jml. *Topic Signature*	Kalimat Terpilih
1 s.d. 3	Sekitar 1 dari 5 orang yang terinfeksi virus Zika menjadi sakit, penyakit Zika (zika disease) atau demam Zika (zika fever). (Kalimat A)
4 s.d. 17	Penderita kanker limfoma atau kanker sel darah putih, kanker kerongkongan atau kanker esofagus, kanker usus, kanker prostat, kanker serviks, dan kanker paru-paru memiliki risiko lebih tinggi terkena kanker lambung. (Kalimat B)
18 s.d. 20	Sekitar 1 dari 5 orang yang terinfeksi virus Zika menjadi sakit, penyakit Zika (zika disease) atau demam Zika (zika fever). (Kalimat A)

Hasil pemilihan kalimat pertama yang dijelaskan pada Tabel 3 menggunakan kata pencarian "gejala yang dirasakan penderita demam zika". Kalimat A merupakan kalimat yang berasal dari dokumen dengan topik penyakit menular. Sedangkan kalimat B bukan berasal dari dokumen dengan topik penyakit menular. Kalimat B terpilih karena mengandung kata "kanker" dalam jumlah yang cukup banyak sementara kata "kanker" merupakan kata yang kurang relevan terhadap topik penyakit menular namun memiliki LLR yang cukup tinggi.

Hasil ini menunjukkan bahwa untuk mendapatkan kalimat pertama ringkasan yang relevan terhadap topik dan sesuai dengan kata pencarian dengan metode ini dibutuhkan *topic signature* yang sesuai dengan topik. Verifikasi manual diperlukan terutama jika *topic signature* didapatkan dengan menggunakan metode otomatis.

C. Pemilihan Kalimat Ringkasan dengan menggunakan MMR

Setelah kalimat pertama terpilih, MMR dapat digunakan untuk mendapatkan kalimat selanjutnya. MMR akan memilih kalimat yang sesuai dengan kata pencarian namun memiliki informasi yang berbeda dari kalimat yang sudah terpilih sebelumnya. Preferensi pada metode pemilihan ini dilakukan dengan mengubah nilai pada MMR dan batas minimal kata dalam kalimat.

Tabel 4 merupakan contoh hasil ringkasan yang didapat dari proses peringkasan dengan menggunakan metode yang diusulkan.

Tabel 4 Contoh Hasil Ringkasan

(A) Sekitar 1 dari 5 orang yang terinfeksi virus Zika menjadi sakit, penyakit Zika (zika disease) atau demam Zika (zika fever). Gejala ini umumnya muncul tidak lama setelah gejala demam mulai dirasakan. Gejala yang paling umum dari Zika adalah demam, ruam, nyeri sendi, atau konjungtivitis (mata merah). demam (demam ringan sampai demam tinggi). Periksa ke dokter jika Anda mengalami gejala yang dijelaskan di atas dan telah mengunjungi daerah di mana Zika ditemukan.
(B) Sekitar 1 dari 5 orang yang terinfeksi virus Zika menjadi sakit, penyakit Zika (zika disease) atau demam Zika (zika fever). Gejala ini umumnya muncul tidak lama setelah gejala demam mulai dirasakan. Demam. Gejala yang paling umum dari Zika adalah demam, ruam, nyeri sendi, atau konjungtivitis (mata merah). Gejala Virus Zika. Gejala Zika mirip dengan demam berdarah dan chikungunya, penyakit menyebar melalui nyamuk yang sama dengan yang menularkan Zika yaitu nyamuk Aedes.

Ringkasan (A) didapat dengan menggunakan kata pencarian "gejala yang dirasakan penderita demam zika", nilai , menggunakan 3 *topic signature*, dan kalimat minimal mengandung 3 kata. Ringkasan (A) merupakan salah satu ringkasan yang cukup baik (berdasarkan pengamatan manual) karena memberikan informasi yang cukup dan sesuai dengan kata pencarian, meskipun belum membentuk ringkasan yang koheren.

Ringkasan (B) merupakan contoh kondisi dimana terjadi kemunculan kalimat palsu akibat tidak digunakannya batas minimal kata dalam kalimat. Ringkasan B didapat dengan kata pencarian yang sama, ,

menggunakan 3 *topic signature*, tanpa batas minimal kata dalam kalimat.

Untuk memperbaiki permasalahan koherensi yang terjadi pada Ringkasan (A) dilakukan penyusunan ulang urutan kalimat dalam ringkasan. Penyusunan urutan kalimat dilakukan dengan memanfaatkan lokasi kalimat pada dokumen asal. Dengan mengurutkan nilai hasil bagi indeks kalimat dengan jumlah kalimat pada dokumen asal, ringkasan dapat disusun ulang urutannya dengan tujuan meningkatkan koherensi.

$$skor\_posisi = \frac{indeks\_kalimat}{jumlah\_kalimat} \quad (10)$$

Kalimat diurutkan kembali berdasarkan nilai pada persamaan (10). Tabel 5 memuat Ringkasan (A) pada Tabel 4 yang telah disusun ulang urutan kalimatnya.

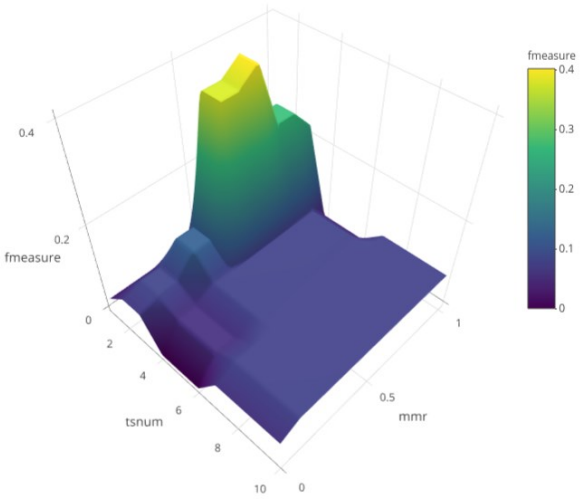
Tabel 5 Ringkasan yang telah diurutkan

(A) Sekitar 1 dari 5 orang yang terinfeksi virus Zika menjadi sakit, penyakit Zika (zika disease) atau demam Zika (zika fever). Gejala yang paling umum dari Zika adalah demam, ruam, nyeri sendi, atau konjungtivitis (mata merah). Periksa ke dokter jika Anda mengalami gejala yang dijelaskan di atas dan telah mengunjungi daerah di mana Zika ditemukan. demam (demam ringan sampai demam tinggi). Gejala ini umumnya muncul tidak lama setelah gejala demam mulai dirasakan.

D. Evaluasi Metode Peringkasan

Evaluasi metode peringkasan secara keseluruhan dilakukan dengan melakukan evaluasi ROUGE terhadap hasil ringkasan yang dihasilkan oleh metode peringkasan otomatis. Evaluasi dilakukan dengan menggunakan ROUGE-S9 dengan stop word removal yang cukup baik digunakan untuk evaluasi peringkasan multi dokumen [19]. Peringkasan dilakukan dengan kombinasi nilai MMR dan jumlah *topic signature* yang digunakan dalam proses peringkasan. 110 hasil ringkasan didapatkan dengan menggunakan nilai ,dan sebagai jumlah *topic signature* yang digunakan. Gambar 3 merupakan nilai akurasi ROUGE-S9 terhadap nilai dan jumlah *topic signature*.

Nilai akurasi tertinggi didapat pada proses peringkasan dengan dengan menggunakan *topic signature* sebanyak 2 dan 3 kata. Kondisi dengan nilai akurasi terendah didapat dengan dengan *topic signature* sebanyak 4, 5, 6. Nilai akurasi yang relatif lebih rendah daripada hasil lainnya pada kondisi ringkasan yang dihasilkan dengan *topic signature* sebanyak 4, 5, 6, membuktikan pengaruh *topic signature* dalam menghasilkan ringkasan yang baik.



Gambar 3 F-Measure ROUGE-S9

V. KESIMPULAN

Dalam penelitian ini, diusulkan peringkasan otomatis multi dokumen untuk Bahasa Indonesia dengan memanfaatkan Log Likelihood Ratio dan *Maximal Marginal Relevance* (MMR) untuk menghasilkan ringkasan yang relevan dan minim redundansi informasi. Dari implementasi dan hasil yang didapatkan selama penelitian dapat ditarik beberapa kesimpulan.`

Pemilihan *topic signature* dengan menggunakan Log Likelihood Ratio (LLR) sangat bermanfaat untuk topik dan bahasa dimana *topic signature* belum tersedia. Namun penggunaan cut-off nilai LLR sebagai kriteria pemilihan tidak sesuai dalam kasus topik yang digunakan dalam penelitian ini (topik mengenai penyakit menular).

Metode peringkasan otomatis ekstraktif, meskipun lebih mungkin diterapkan dalam berbagai kasus, memiliki kelemahan dalam pembentukan ringkasan yang koheren.

Topic signature (dan akurasinya) sangat mempengaruhi hasil peringkasan otomatis dengan metode yang diusulkan pada penelitian ini. Dalam penelitian ini, satu *topic signature* yang tidak sesuai dapat menyebabkan akurasi hasil ringkasan jatuh dibawah 10% meskipun menggunakan nilai MMR yang sama dengan hasil ringkasan dengan akurasi tertinggi.

Metode peringkasan yang diusulkan memberikan nilai akurasi maksimal sebesar 0,4 (diukur dengan menggunakan ROUGE-S9). Meskipun nilai ini dirasa belum memuaskan, membandingkan nilai akurasi metode peringkasan dokumen otomatis Bahasa Indonesia tidak mudah dilakukan karena perbedaan metode evaluasi dan tidak tersedianya dataset terstandar untuk Bahasa Indonesia.

Dari penelitian ini, masih terdapat beberapa potensi penelitian lanjutan yang dapat dilakukan untuk peringkasan otomatis Bahasa Indonesia. Pemilihan *topic signature* secara otomatis perlu diteliti lebih lanjut mengingat pengaruh yang diberikan terhadap hasil akhir ringkasan. Selain itu, *topic signature* juga dapat digunakan diluar ranah penelitian peringkasan otomatis. Tidak tersedianya dataset merupakan salah satu kendala yang dihadapi dalam penelitian ini. Pembuatan dataset terstandar seperti DUC2001-2007 [23] atau MultiLing [24] sangat diperlukan untuk kemajuan penelitian peringkasan otomatis di Indonesia.

REFERENSI

[1] S. Kemp, "Digital, Social & Mobile in APAC in 2015," 2015.

[2] S. Kemp, "Digital in 2016," 2016.

[3] We Are Social and Hootsuite, "Digital in 2017: Global Overview," 2017.

[4] T. O'Reilly, "What Is Web 2.0 - O'Reilly Media," *OReillycom*, 2005. [Online]. Available: http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html. [Accessed: 01-Sep-2016].

[5] D. Bawden and L. Robinson, "The dark side of information: overload, anxiety and other paradoxes and pathologies," *J. Inf. Sci.*, vol. 35, no. 2, pp. 180–191, Apr. 2009.

[6] N. Hayatin, G. I. Marthasari, N. Hayatin, "Improvement of sentences scoring based news feature for news summary on social media issues," *Semin. Nas. Teknol. dan Rekayasa*, pp. 1–5, 2017.

[7] S. Verdianto, A. Z. Arifin, and D. Purwitasari, "Strategi pemilihan kalimat pada peringkasan multi dokumen," *Nusant. J. Comput. its Appl.*, vol. 2, no. 7, pp. 1–5, 2016.

[8] N. Hayatin, C. Fatichah, and D. Purwitasari, "Pembobotan kalimat berdasarkan fitur berita dan trending issue untuk peringkasan multidokumen berita," *JUTI J. Ilm. Teknol. Inf.*, vol. 13, no. 1, p. 38, Jan. 2015.

[9] H. P. Luhn, "The Automatic Creation of Literature Abstracts," *IBM J. Res. Dev.*, vol. 2, no. 2, pp. 159–165, 1958.

[10] H. P. Edmundson, "New methods in automatic extracting," *J. Assoc. Comput. Mach.*, vol. 16, no. 2, pp. 264–285, 1969.

[11] P. B. Baxendale, "Machine-Made Index for Technical Literature: An Experiment," *IBM J. Res. Dev.*, vol. 2, no. 4, pp. 354–361, 1958.

[12] I. Mani and M. T. Maybury, *Advances in Automatic Text Summarization*, vol. 26, no. 2. 1999.

[13] C. Lin, "Training a selection function for extraction," *CIKM '99 Proc. eighth Int. Conf. Inf. Knowl. Manag.*, pp. 55–62, 1999.

[14] D. Das and A. F. T. Martins, "A Survey on Automatic Text Summarization," *Eighth ACIS Int. Conf. Softw. Eng. Artif. Intell. Netw. ParallelDistributed Comput. SNPD 2007*, vol. 4, pp. 574–578, 2007.

[15] K. Mckeown and D. R.Radev, "Generating Summaries of Multiple News Articles," *Proc. 18th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 74–82, 1995.

[16] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 335–336, 1998.

[17] D. R. Radev, H. Jing, and M. Budzikowska, "Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies," *Inf. Process. Manag. 40.6 919-938.*, vol. 40, no. 6, p. 10, 2000.

[18] S. Gupta, A. Nenkova, and D. Jurafsky, "Measuring importance and query relevance in topic-focused multi-document summarization," *Proc. 45th Annu. Meet. ACL Interact. Poster Demonstr. Sess. - ACL '07*, no. C, p. 193, 2007.

[19] C. Lin and E. Hovy, "The automated acquisition of topic signatures for text summarization," *Proc. 18th Conf. Comput. Linguist. -*, vol. 1, pp. 495–501, 2000.

[20] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," *M.Sc. Thesis, Append. D*, vol. pp, pp. 39–46, 2003.

[21] A. Mirna, A. Jelita, N. Bobby, S. M. M. Tahaghoghi, and E. W. Hugh, "Stemming Indonesian: A confix-stripping approach," *ACM Trans. Asian Lang. Inf. Process.*, vol. 6, no. 4, pp. 1–33, 2007.

[22] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Proc. Work. text Summ. branches out (WAS 2004)*, no. 1, pp. 25–26, 2004.

[23] NIST, "Document Understanding Conferences http://duc.nist.gov/," 2007. [Online]. Available: http://www-nlpir.nist.gov/projects/duc/data/2007_data.html. [Accessed: 16-Aug-2017].

[24] NIST, "MultiLing Pilot - Multi - Lingual Summarization in the Text Analysis Conference (TAC) 2011." [Online]. Available: http://users.iit.demokritos.gr/~ggianna/TAC2011/MultiLing2011.html. [Accessed: 16-Aug-2017].