Jurnal Ilmiah Teknik Informatika dan Komunikasi, Vol.1, No.3, November 2021, pp. 77-88
p-ISSN : 2827-8127 (print)
e-ISSN : 2827-7945 (online)
http://journal.sinov.id/index.php/JUITIK

MAXIMUM MARGINAL RELEVANCE BERBASIS *BOOLEAN MODEL* PADA PERINGKASAN ARTIKEL BERITA PENDEK

Arie Atwa Magriyanti
arie.atwa@stekom.ac.id

Universitas Sains dan Teknologi Komputer
Semarang

Abstrak

Portal berita *online* merupakan situs yang memuat segala berita dan artikel untuk dibaca pengunjung, berisi opini dan komentar-komentar seputar politik, teknologi dan lain-lain. Ada berbagai portal berita *online* yang bisa kita akses, antara lain antaranews.com dan tribunnews.com. Meningkatnya jumlah portal berita *online,* mengakibatkan tingginya jumlah berita yang bisa dibaca masyarakat. Oleh sebab itu, kebutuhan peringkasan teks *(text summarization)* semakin diperlukan masyarakat untuk kemudahan dan penghematan waktu. Sistem peringkasan teks otomatis yang digunakan pada penelitian sebelumnya, menggunakan metode *Maximum Marginal Relevance* (MMR) berbasis *Vector Space Model* (VSM) dengan representasi algoritma pembobotan kata menggunakan TF-IDF-DF *(Term Frequency-Inverse Document Frequency-Document Frequency)*. Pada algoritma ini, kata yang sering muncul memiliki jumlah yang tinggi, sehingga bobot hubungan antara sebuah kata dan kalimat rendah, sehingga metode ini cocok untuk artikel panjang yang memiliki banyak jumlah kalimat. Oleh karena itu, perlu diusulkan metode lain untuk menghitung kesamaan kata dengan menggunakan *boolean model* dengan representasi *jaccard, dice* dan *cosine coefficient.* Ketiga metode ini digunakan untuk mengetahui *document* yang paling relevan untuk kumpulan kata kunci *(query)* yang diberikan. Setelah proses *boolean*, perlu dilakukan metode ekstraksi teks yang diterapkan yaitu MMR *(Maximum Marginal Relevance)* untuk meringkas *document* tunggal dengan cara melakukan rangking, membandingkan similarity *query* dan *document*, dan *similarity* antar *document*. Dari hasil penelitian, *boolean model* memiliki nilai akurasi yang lebih tinggi daripada VSM, dan di antara ketiga metode *boolean model*, metode *cosine coefficient* lebih unggul dengan akurasi 59.3 %.

Kata kunci : *text summarization*, VSM, TF-IDF-DF, *boolean model, jaccard, dice, cosine coefficient*, MMR

ABSTRACT

Online news portal is a site that contains all news and articles to read visitors, contains opinions and comments about politics, technology and others. There are various online news portals that we can access, including antaranews.com and tribunnews.com. The increasing number of online news portals, resulting in a high number of news that can be read by the public. Therefore, the need for text summarization is increasingly needed by the community for ease and time savings. The automated text summary system used in previous research uses the Maximum Marginal Relevance (MMR) method based on Vector Space Model (VSM) with the word weighting algorithm representation using TF-IDF-DF (Term Frequency-Inverse Document Frequency-Document Frequency). In this algorithm, words that often appear have a high number, so the weight of the relationship between a word and sentence is low, so this method is suitable for long articles that have many sentences. Therefore, it is necessary to propose another method to compute word equality by using boolean model with representation of

Received Agustus 19, 2021*; Revised September 01,* 2021*; Accepted Oktober 25,* 2021

jaccard, dice and cosine coefficient. These three methods are used to find the most relevant documents for the given set of queries. After the boolean process, a text extraction method that is applied is MMR (Maximum Marginal Relevance) to summarize a single document by ranking, comparing similarity queries and documents, and similarity between documents. From the research results, boolean models have higher accuracy values than VSM, and among the three boolean model methods, the cosine coefficient method is superior with accuracy 59.3 %.

Keyword : text summarization, VSM, TF-IDF-DF, *boolean model, jaccard, dice, cosine coefficient*, MMR

A. PENDAHULUAN

Di era globalisasi, teknologi informasi dan komunikasi berkembang pesat. Informasi merupakan hal yang dibutuhkan seluruh masyarakat baik dari berbagai umur maupun kalangan dengan tujuan untuk menambah pengetahuan. Demi memenuhi kebutuhan informasi, masyarakat menggunakan berbagai media dan cara yang ada. Untuk memperoleh informasi telah menggunakan media yang lebih efektif dan mudah, baik informasi tentang berita-berita teraktual antara lain berita politik, kriminalitas, olahraga, fashion, maupun gaya hidup dan lain-lain. Media ini disebut dengan internet atau media *online* yang menjadi alat komunikasi penting dan dibutuhkan banyak orang. Media *online* menjadi kebutuhan primer dan populer di masyarakat.

Dalam penggunaan media *online*, ada berbagai jenis *website* yang bisa ditemukan, salah satunya adalah portal berita. Portal berita *online* merupakan situs yang memuat segala berita dan artikel untuk dibaca pengunjung, berisi opini dan komentar-komentar seputar politik, teknologi dan lain-lain. Berbagai situs berita ini seperti layaknya koran, buletin dan majalah.

Ada berbagai portal berita *online* yang bisa diakses, antara lain antaranews.com dan tribunnews.com. Portal ini menyediakan berbagai macam berita yang terjadi baik berskala nasional maupun lokal secara aktual dan cepat. Meningkatnya jumlah portal berita *online,* mengakibatkan tingginya jumlah berita yang bisa dibaca masyarakat. Oleh sebab itu, kebutuhan peringkasan teks semakin diperlukan masyarakat untuk kemudahan dan penghematan waktu.

Penelitian pada jurnal[1] disebutkan bahwa sistem peringkasan teks otomatis menggunakan metode *Maximum Marginal Relevance* (MMR) berbasis algoritma pembobotan kata menggunakan algoritma TF-IDF-DF. Metode TF-IDF-DF *(Term Frequency-Inverse Document Frequency-Document Frequency)* merupakan modifikasi dari metode TF-IDF untuk mendapatkan bobot perwakilan dari kata-kata yang diekstrak dari data informasi dengan mempertimbangan penyebaran kata di *document* lain. Sedangkan MMR adalah metode untuk menentukan relevansi hasil ringkasan dengan *document* dan *query* (judul artikel berita) yang diberikan oleh *user* berdasarkan bobot kesamaannya serta dapat mengurangi redundansi dalam peringkasan.

Metode TF-IDF-DF adalah metode pembobotan yang menggunakan konsep *term frequency* yaitu berdasar pada kata yang sering muncul. Metode TF-IDF-DF sendiri merupakan representasi dari *Vector Space Model* (VSM). *Document* dalam VSM berupa matriks yang berisi bobot seluruh kata pada tiap *document*. Bobot tersebut menyatakan kepentingan atau kontribusi kata terhadap suatu *document* dan kumpulan *document*. Kepentingan suatu kata dalam *document* dapat dilihat dari frekuensi kemunculannya terhadap *document*. Jumlah kalimat yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut. Bobot hubungan antara sebuah kata dan sebuah kalimat akan tinggi apabila jumlah kata tersebut tinggi di dalam kalimat dan jumlah keseluruhan kalimat yang mengandung kata tersebut rendah pada kumpulan *document (database).*

Berdasarkan hasil pembobotan kata di penelitian[1] menggunakan metode TF-IDF-DF yang merupakan artikel berita pendek *(short article)* berbahasa Indonesia, diperoleh bahwa jumlah keseluruhan kalimat yang mengandung kata yang sering muncul memiliki jumlah yang tinggi, sehingga bobot hubungan antara sebuah kata dan kalimat rendah, padahal kata yang sering muncul tersebut menjadi topik pembicaraan dari isi berita, atau dengan kata lain menjadi kata penting. Dan jika dalam satu artikel jumlah seluruh kalimat (N) sama dengan *document frequency* yang

merupakan banyaknya kalimat yang berisi kata ke-i *(df_i)*, maka pembobotan akan bernilai nol (0). Dengan demikian, metode ini cocok untuk artikel panjang yang memiliki banyak jumlah kalimat.

Berdasarkan analisa di atas, diperoleh bahwa metode *term weighting* menggunakan VSM yang dipresentasikan dengan metode TF-IDF-DF, jika digunakan untuk melakukan pembobotan kata pada artikel pendek berbahasa Indonesia ternyata kurang tepat. Oleh karena itu, perlu diusulkan metode lain untuk menghitung kesamaan kata *(similarity)* dengan tidak menggunakan model pembobotan, tetapi berdasarkan *distance measure* menggunakan *boolean model* yang dipresentasikan melalui metode *jaccard, dice,* dan *cosine coefficient.*

Pada proses penelusuran *boolean model*, setiap kata diubah ke dalam ekspresi *boolean* sehingga menghasilkan sebuah aturan *binary*. Variabel nilai bobot selalu bersifat biner (dua pilihan), yaitu nol atau satu. Jika nilainya satu maka model *boolean* menyimpulkan bahwa *document* relevan terhadap sebuah permintaan *(query)*. Selebihnya, kalau bernilai nol maka *document* dianggap tidak relevan.

Penelitian pada jurnal[2] menyebutkan tiga kesamaan koefisien yaitu *jaccard, dice dan cosine coefficient* digunakan untuk mengetahui *document* yang paling relevan untuk kumpulan kata kunci *(query)* yang diberikan. Selanjutnya perlu dilakukan metode ekstraksi teks yang diterapkan untuk meringkas *document* tunggal dengan cara melakukan rangking dan membandingkan kesamaan *(similarity)* antar *document*. Metode yang digunakan adalah *Maximum Marginal Relevance (MMR)*. Tujuan MMR ini untuk memperoleh skor kalimat berdasarkan *similarity* kalimat dengan *query* yang diberikan.

Berdasarkan pemaparan di atas, maka penelitian yang akan dilakukan penulis adalah melakukan proses peringkasan pada dataset [1] yaitu berita pendek berskala nasional dan lokal (artikel berita pendek) berbahasa Indonesia menggunakan *Maximum Marginal Relevance (MMR)* berbasis *boolean model (jaccard, dice dan cosine coefficient).*

B. LANDASAN TEORI

1. Konsep Dasar Peringkasan Teks Otomatis

Dalam dunia komputer peringkasan teks dikenal dengan *Automatic Summarization Text* atau peringkasan teks otomatis. Pada penelitian ini, penulis memilih metode *Maximum Marginal Relevance* (MMR) untuk meringkas artikel berita pendek dengan metode perangkingan. Menurut www.romelteamedia.com[17] dan observasi penulis pada dataset (lampiran 1) sebuah artikel berita dikatakan sebagai artikel berita pendek jika memiliki panjang kata maksimal 800 kata. Hal ini dimaksudkan supaya sesuai karakter bahasa jurnalistik yaitu lugas, ringkas, sederhana, dan mudah dipahami.

2. Text Preprocessing

Text preprocessing adalah suatu tahap untuk mengolah teks berita yang merupakan bahan mentah menjadi kata-kata yang telah siap dihitung bobot katanya. Beberapa proses dari *text preprocessing*, yaitu segmentasi kalimat, *case folding, tokenizing, filtering,* dan *stemming.*

3. Konsep Dasar VSM dengan Representasi TF-IDF-DF

TFIDF adalah metode pembobotan paling umum yang digunakan untuk menggambarkan *document* dalam *Vector Space Model* (VSM), terutama dalam masalah *Information Retrieval*. Metode TF-IDF-DF merupakan modifikasi dari metode TF-IDF, karena metode TF-IDF memiliki kekurangan dalam pembobotan kata. Kekurangannya yaitu adanya anggapan bahwa kata yang tersebar dalam *document* lain tidak penting, sehingga dianggap tidak ada. Padahal kata yang sering muncul dalam kalimat lain bisa jadi merupakan kata yang penting. Akibatnya, nilai bobot yang tinggi diperoleh pada kata yang memiliki frekuensi tinggi dalam *document*, sedangkan kata yang tersebar di *document* lain memiliki perhitungan bobot yang kecil.

4. Konsep Dasar *Boolean Model* dengan Representasi *Jaccard, Dice,* dan *Cosine Coefficient)*

Document merupakan himpunan dari istilah *(term)* sedangkan query merupakan pernyataan *boolean* yang ditulis pada *term*. *Term* dalam sebuah *query* dihubungkan dengan menggunakan operator AND, OR atau NOT. Pada *boolean model* setiap kata diubah ke dalam ekspresi *boolean* sehingga menghasilkan sebuah aturan *binary*. Variabel nilai bobot selalu bersifat biner (dua pilihan), yaitu nol atau satu. Pada peneltian ini dilakukan uji pengukuran kemiripan *document* dengan menggunakan rumus *jaccard, dice,* dan *cosine coefficient*. Ketiga pengukuran ini merupakan pengukuran yang terbaik dari beberapa pengukuran yang ada.[7]

5. Konsep Dasar *Maximum Marginal Relevance* (MMR)

Summarization bertujuan untuk menghasilkan ringkasan sebuah *document* atau sekelompok *document*. *Text summarization* dapat dikategorikan dalam peringkasan *single-document* atau *multi-document*. Peringkasan pada *single-document* mengusulkan metode *Maximum Marginal Relevance* (MMR) untuk menghasilkan ringkasan. Metode ini diusulkan pertama kali oleh Carbonell dan Goldstein pada tahun 1998.[10]

6. Evaluasi Peringkasan Teks

Menurut Nedunchelian[15], proses evaluasi hasil *text summarization* dilakukan menggunakan tiga parameter yaitu *precision, recall,* dan *F-measure.*

C. METODE PENELITIAN

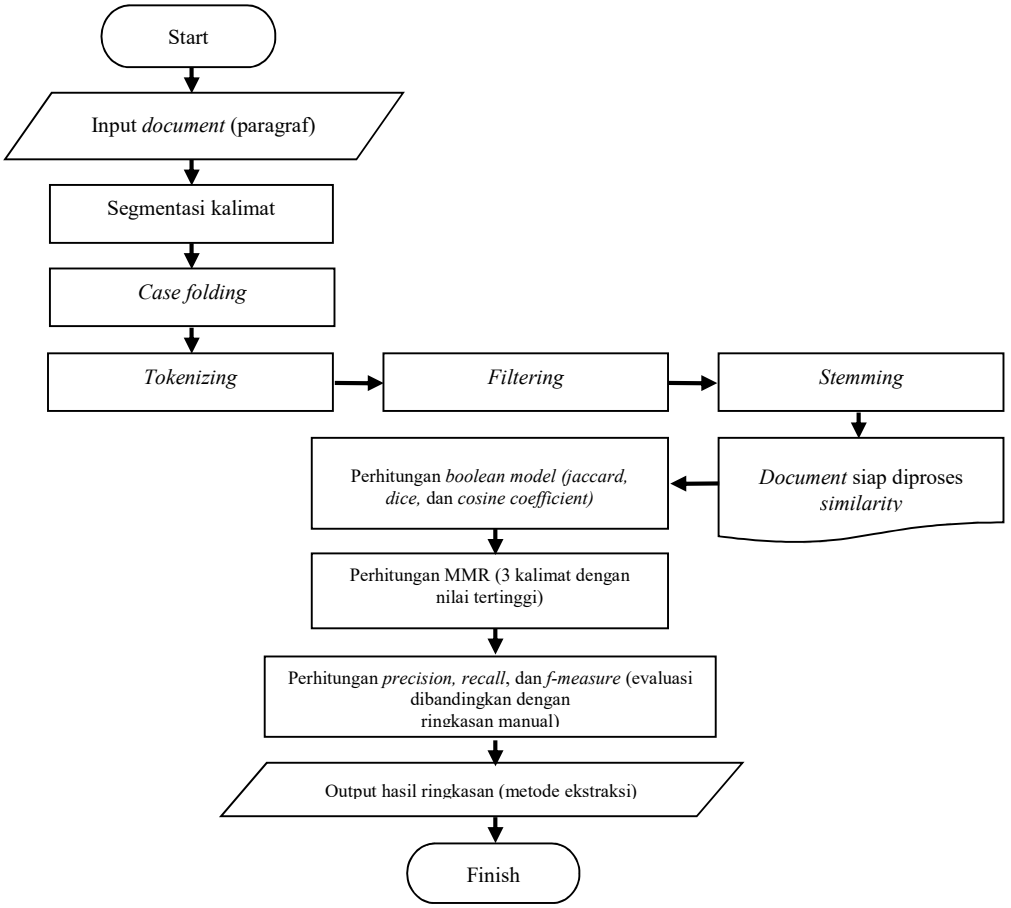
1. Metode Pengumpulan Data

Metode pengumpulan data yang diterapkan dalam memperoleh data yang dibutuhkan yaitu observasi dan studi pustaka.

2. Eksperimen

Proses eksperimen yang dilakukan pada penelitian ini ditunjukkan oleh alur di bawah ini:

Gambar 1. Alur Proses Eksperimen



3. Evaluasi

Untuk mengetahui kualitas ringkasan dengan sistem peringkasan teks, maka hasil tersebut akan dibandingkan dengan ringkasan manual. Untuk peringkasan manual ini, para responden meringkas teks berita terlebih dahulu, dengan langkah:

a. Dari 30 judul tersebut, 16 judul hasil peringkasan manual mengacu pada penelitian[1] dan 14 judul lainnya menunjuk tiga orang sebagai responden (R1, R2, R3). Ketiga responden ini adalah guru Bahasa Indonesia SMK Negeri 11 Semarang yang bernama Sri Nurhidayati, S.Pd., Alfiyah, S.Pd., dan Slamet, S.Pd. Masing-masing responden bertugas meringkas 7 judul berita berskala nasional dan 7 judul berita berskala lokal.
b. Tiga guru tersebut memilih 3 kalimat yang paling sesuai dengan judul berita sebagai kalimat hasil ringkasan.

Dari hasil ringkasan yang dilakukan para responden dan sistem komputasi peringkasan teks menggunakan *boolean model (jaccard, dice, cosine coefficient)* dan MMR, akan dihitung tingkat akurasinya dengan menggunakan tiga parameter, yaitu *precision, recall,* dan *f-measure.* Dari hasil ini maka penelitian bisa dihitung apakah keakuratan meningkat atau tidak.

D. HASIL PENELITIAN DAN PEMBAHASAN

1. Kebutuhan Data Teks

Data teks berita online terdiri dari 15 judul berita berskala nasional (artikel 1 s.d 15) dan 15 judul berita berskala lokal (artikel 16 s.d. 30)

2. *Text Preprocessing*

Untuk mengolah teks berita menjadi kata-kata yang siap dihitung pada proses selanjutnya perlu dilakukan *preprocessing* terlebih dahulu. Beberapa proses *text preprocessing*, yaitu segmentasi kalimat, *case folding, tokenizing, filtering,* dan *stemming.*

3. Perhitungan *Boolean Model*

Setelah *text preprocessing*, tahap selanjutnya adalah proses menghitung tingkat kemiripan *document*. Pada *boolean model* setiap kata diubah ke dalam ekspresi *boolean* sehingga menghasilkan sebuah aturan *binary*. Variabel nilai bobot selalu bersifat biner (dua pilihan), yaitu nol (0) atau satu (1). *Boolean model* yang diterapkan pada penelitian ini antara lain metode *jaccard, dice, dan cosine coefficient.*

Jaccard Coefficient merupakan metode yang digunakan untuk membandingkan kesamaan dan keragaman 2 set sampel dengan rumus:

$$Jaccard\ (Q,D) = \frac{|Q \cap D|}{|Q \cup D|} = \frac{|Q \cap D|}{|Q| + |D| - |Q \cap D|}$$

Pada *dice coefficient*, kemiripan diperoleh dengan menghitung 2 kali jumlah atribut yang sama yang dimiliki oleh kedua *document Q* dan *D*, setelah itu dibagi dengan jumlah kata *document Q* yang tidak dimiliki oleh *document D* ditambah jumlah kata *document D* yang tidak dimiliki oleh *document Q*. Berikut rumus *dice coefficient* dan untuk keterangan notasi *dice* sama dengan notasi *jaccard.*

$$Dice\ (Q,D) = \frac{\mathbf{2}|\boldsymbol{Q} \cap \boldsymbol{D}|}{|\boldsymbol{Q}| + |\boldsymbol{D}|}$$

Cosine coefficient merupakan metode ukuran kesamaan dengan menghitung sudut antara vektor *document* dengan vektor kueri. Vektor ini diartikan sebagai satuan panjang, *cosinus* dari sudut antara mereka hanyalah *dot product* dari vektor. Berikut rumus untuk menerapkan *cosine coefficient,* untuk notasi keterangan *cosine* juga sama dengan *jaccard.*

$$Cosinus\ (Q,D) = \frac{|Q \cap D|}{|Q|^{1/2} * |D|^{1/2}}$$

Menghitung tingkat kemiripan *document* dibagi menjadi 2 tahap, yaitu :

1) Perhitungan relevansi antara *document* dan *query* (judul)
2) Perhitungan *similarity* antara *document*

4. Perhitungan MMR

Algoritma MMR digunakan untuk merangking kalimat-kalimat sebagai tanggapan terhadap *query* yang diberikan *user.* Prinsip perhitungan metode MMR yaitu mengambil kalimat dengan nilai tertinggi dari setiap perhitungan iterasi. Adapun nilai parameter λ yang digunakan adalah 0,85. Proses perhitungan MMR sebagai berikut dengan catatan *Sim1(Di, Q)* adalah *relevance query.* Sedangkan *Sim1(Di, Dj)* adalah *similarity* kalimat terhadap kalimat yang diekstrak :

$$MMR\ (D_i) = \lambda\ .\ Sim_1\ (D_i, Q) - (1 - \lambda).max\ Sim_2\ (D_i, D_j)$$

Untuk peringkasan secara sistem komputasi, penulis menggunakan program PyCharm. Pada program PyCharm ini, dilakukan proses peringkasan untuk masing-masing *boolean model* yaitu *jaccard*, *dice*, dan *cosine coefficient* dan sekaligus proses MMR (merangking menjadi 3 kalimat hasil ringkasan). Tabel berikut adalah hasil ringkasan artikel berita nasional dan lokal menggunakan program PyCham.

Tabel 1. Hasil Ringkasan Menggunakan Program PyCharm

ARTIKEL		HASIL RINGKASAN SISTEM		
		JACCARD	DICE	COSINE
		Document ke-	Document ke-	Document ke-
Berita Nasional	1	0, 6, 2	0, 6, 2	0, 6, 2
	2	0, 3, 2	0, 3, 2	0, 3, 2
	3	2, 0, 7	2, 0, 7	2, 0, 7
	4	3, 0, 2	3, 0, 2	3, 0, 2
	5	0, 5, 2	0, 5, 2	0, 5, 2
	6	0, 8, 1	0, 8, 1	0, 8, 1
	7	0, 2, 1	0, 2, 1	0, 2, 1
	8	0, 2, 5	0, 2, 5	0, 2, 1
	9	0, 2, 3	0, 2, 11	0, 2, 3
	10	0, 2, 3	0, 2, 3	0, 2, 1
	11	0, 5, 3	0, 5, 3	0, 5, 1
	12	1, 0, 10	1, 0, 10	1, 0, 10
	13	0, 1, 3	0, 3, 1	0, 1, 3
	14	0, 13, 1	0, 10, 1	0, 10, 1
	15	2, 0, 3	2, 0, 3	2, 0, 3
Berita Lokal	16	0, 10, 13	0, 10, 13	0, 10, 13
	17	4, 0, 3	4, 0, 3	4, 0, 8
	18	0, 10, 2	0, 10, 2	0, 2, 10
	19	4, 1, 0	4, 1, 0	4, 1, 0
	20	0, 2, 1	0, 2, 1	0, 2, 1
	21	2, 1, 6	2, 1, 6	2, 1, 6
	22	0, 1, 5	0, 1, 5	0, 1, 5
	23	0, 1, 8	0, 1, 8	0, 1, 8
	24	5, 0, 9	5, 0, 9	5, 0, 9
	25	2, 9, 10	2, 9, 10	2, 9, 10
	26	0, 3, 1	0, 3, 1	0, 3, 1
	27	0, 7, 8	0, 7, 8	0, 7, 8
	28	11, 0, 1	11, 0, 1	11, 0, 2
	29	1, 2, 4	1, 2, 4	1, 2, 4
	30	1, 0, 2	1, 0, 2	1, 0, 2

Peringkasan manual diperlukan untuk proses evaluasi. Oleh karena itu, penulis mengambil 3 responden untuk melakukan peringkasan manual ini. Dari 30 jumlah artikel, responden hanya meringkas 14 artikel (artikel nomor 9-15 dan artikel nomor 24-30) ditambah 16 artikel berdasarkan hasil ringkasan responden pada penelitian[1] (artikel nomor 1-8 dan artikel nomor 16-23). Hasil ringkasan manual ditunjukkan pada tabel berikut:

Tabel 2. Hasil Ringkasan Manual oleh Ahli Bahasa

ARTIKEL		HASIL RINGKASAN MANUAL		
		Responden 1 (R1)	Responden 2 (R2)	Responden 3 (R3)
		Document ke-	Document ke-	Document ke-
Berita Nasional	1	0, 2, 6	0, 1, 2	0, 1, 4
	2	0, 3	0, 1, 4	0, 2, 3
	3	0, 2, 3	0, 3, 2	0, 2, 3
	4	0, 1, 3	0, 3, 4	0, 3, 4
	5	0, 1, 3	0, 1, 2	0, 1, 2
	6	0, 1, 8	0, 1, 2	0, 1, 2
	7	0, 2, 3	0, 2, 3	0, 4, 6
	8	0, 3, 5	0, 3, 5	0, 2, 3
	9	0, 1, 2	1, 3, 7	0, 2, 3
	10	0, 1, 2	2, 3, 4	0, 1, 4
	11	0, 1, 3	0, 1, 4	0, 1, 4
	12	0, 1, 4	0, 4, 6	0, 4, 6
	13	0, 1, 4	0, 2, 4	0, 1, 3
	14	0, 2, 3	0, 1, 2	0, 10, 12
	15	0, 3, 9	0, 1, 3	0, 3
Berita Lokal	16	6, 2, 4	0, 2, 3	0, 3, 8
	17	0, 1, 8	0, 1, 3	0, 1, 4
	18	0, 2, 7	0, 2, 9	0, 2, 5
	19	0, 4, 2	0, 1, 2	2, 6, 7
	20	0, 2, 4	0, 1, 3	0, 1, 2
	21	0, 1, 2	0, 1, 2	0, 2, 5
	22	1, 3, 8	1, 4, 5	0, 1, 3
	23	0, 8, 6	5, 7, 8	7, 5, 8
	24	0, 1, 2	0, 2, 7	0, 2, 3
	25	2, 9, 10	2, 9, 10	2, 9, 10
	26	0, 2, 3	0, 2, 3	0, 2, 3
	27	0, 4, 8	0, 1, 2	0, 1, 6
	28	0, 1, 2	4, 10	0, 2, 10
	29	0, 1, 2	1, 2, 4	0, 2, 4
	30	1, 2, 3	1, 3	1, 2, 3

5. Evaluasi Peringkasan Teks

Evaluasi peringkasan teks oleh sistem terhadap hasil ringkasan manual menggunakan tiga parameter yaitu *precision, recall,* dan *f-measure.*

a. Precision

Merupakan perbandingan jumlah informasi relevan yang didapatkan sistem dengan jumlah seluruh informasi yang terambil oleh sistem baik yang relevan maupun tidak. Persamaan *precision* ditunjukkan pada persamaan berikut:

$$Precision = \frac{correct}{(correct + wrong)}$$

b. Recall

Merupakan perbandingan jumlah informasi relevan yang didapatkan sistem dengan jumlah seluruh informasi relevan yang ada dalam koleksi informasi (baik yang terambil atau tidak terambil oleh sistem).

$$Recall = \frac{correct}{(correct + missed)}$$

Correct : jumlah kalimat yang diekstrak oleh sistem dan manusia.
Wrong : jumlah kalimat yang diekstrak oleh sistem tetapi tidak diekstrak oleh manusia.
Missed : jumlah kalimat yang diekstrak oleh masnusia tetapi tidak diekstrak oleh sistem.

c. *F-measure*

Merupakan hubungan antara *recall* dan *precision* yang mempresentasikan akurasi sistem. Persamaan *F-measure* seperti berikut:

$$F - measure = \frac{2 * Rec * P}{(R + P)}$$

Terlebih dahulu mencari *precision* dan *recall* setiap responden dengan sistem, selanjutnya mencari rata-rata *precision* dan *recall,* untuk menentukan *f-measure* pada artikel berita.

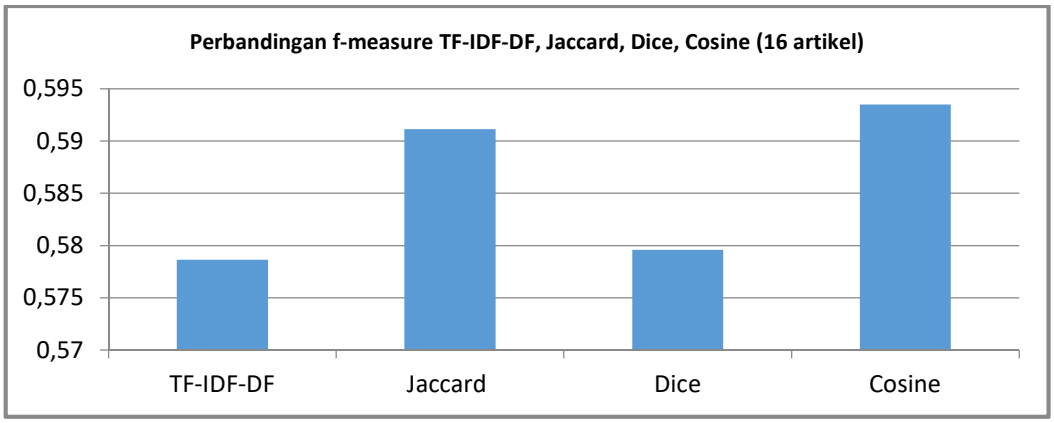
6. Perbandingan Nilai Akurasi antara Metode *Jaccard, Dice,* dan *Cosine Coefficient* dengan Metode TF-IDF-DF

Untuk mengetahui keefektifan metode baru yang diusulkan untuk proses peringkasan artikel berita pendek, maka metode yang digunakan pada penelitian sebelumnya oleh Winda[1] menggunakan metode VSM (TF-IDF-DF) dibandingkan dengan peringkasan teks menggunakan *boolean model* yaitu *jaccard, dice*, dan *cosine coefficient* yang dilakukan pada 16 artikel berita. Berikut adalah perbandingan nilai *f-measure TF-IDF-DF, Jaccard, Dice*, dan *Cosine Coefficient*.

Tabel 3. Perbandingan nilai *f-measure* TF-IDF-DF, *Jaccard, Dice,* dan *Cosine Coefficient*

	Precision	**Recall**	**F-Measure**
TF-IDF-DF	0.597222313	0.555555625	0.578637694
Jaccard	0.590277778	0.592013889	0.591128118
Dice	0.576388889	0.583333333	0.579594017
Cosine	0.590277778	0.597222222	**0.593482906**

Berdasarkan tabel 3 dapat dilihat bahwa metode TF-IDF-DF memiliki akurasi 57.8 %, *jaccard* 59.1 %, *dice* 57.9 %, dan *cosine* 59.3 %. Dengan demikian peringkasan menggunakan *boolean model (jaccard, dice,* dan *cosine)* memiliki nilai akurasi lebih tinggi dibandingkan dengan metode pembobotan menggunakan VSM (TF-IDF-DF). Hasil perbandingan tersebut juga bisa dilihat dari gambar 2 berikut:



Gambar 2. Perbandingan nilai *f-measure* TF-IDF-DF, *Jaccard, Dice,* dan *Cosine* pada 16 artikel

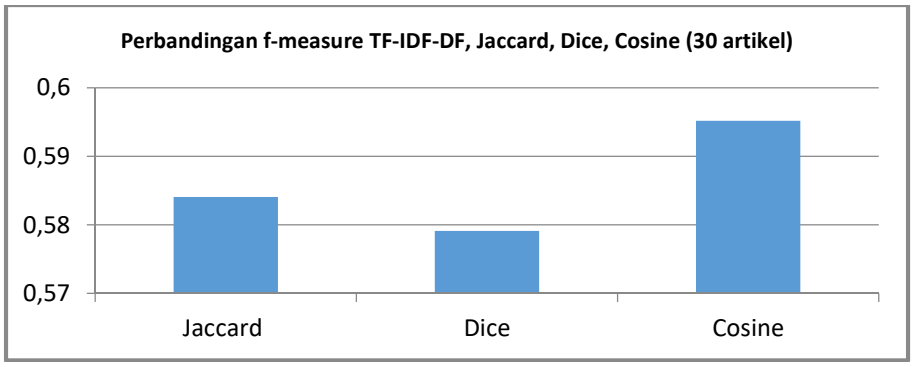
7. Perbandingan Nilai Akurasi antara Metode *Jaccard, Dice*, dan *Cosine Coefficient* pada Seluruh Artikel

Berdasarkan perhitungan peringkasan pada seluruh artikel berita sejumlah 30 artikel, maka dapat dilihat perbandingan nilai akurasi berdasarkan *precision, recall,* dan *f-measure* masing-masing metode *boolean model (jaccard, dice*, *cosine coefficient*). Perbandingan tersebut ditunjukkan pada tabel 4 berikut.

Tabel 4. Perbandingan nilai *f-measure* masing-masing metode *boolean model* pada seluruh artikel berita

	Precision	**Recall**	**F-Measure**
Jaccard	0.581481481	0.587037037	0.584072717
Dice	0.577777778	0.580555556	0.579113127
Cosine	0.592592593	0.598148148	**0.595183829**

Hasil perbandingan tersebut juga bisa dilihat dari diagram batang gambar 3 berikut:



Gambar 3. Perbandingan nilai *f-measure Jaccard, Dice,* dan *Cosine* pada 30 artikel

Berdasarkan tabel 4 dan gambar 3 dapat disimpulkan bahwa pada proses peringkasan 30 artikel berita pendek, metode *cosine* dengan akurasi 59.5 % juga memiliki nilai akurasi lebih tinggi daripada metode *jaccard* dengan akurasi 58.4 % dan *dice* dengan akurasi 57.9%.

E. KESIMPULAN DAN SARAN

1. Kesimpulan

Peringkasan artikel berita pendek menggunakan metode MMR berbasis *boolean model* yaitu *jaccard, dice*, dan *cosine coefficient* dibandingkan dengan penelitian sebelumnya [1] yang menggunakan prinsip *vector space model* (TF-IDF-DF) dapat disimpulkan bahwa *boolean model* memiliki nilai akurasi yang lebih tinggi daripada metode VSM. Di antara ketiga metode *boolean model*, metode *cosine coefficient* lebih unggul yaitu dengan akurasi sebesar 59.3 %. Metode ini tepat digunakan untuk mengetahui *document* yang paling relevan untuk kumpulan kata kunci *(query)* yang diberikan, dibandingkan dengan metode TF-IDF-DF yang perhitungannya menggunakan pembobotan kata yang sering muncul dalam *document,* di mana kata yang sering muncul justru memiliki bobot lebih rendah.

Berikutnya hasil yang diperoleh untuk seluruh artikel berita pendek yang penulis uji cobakan pada 30 artikel, diperoleh bahwa dengan akurasi 59.5 % metode *cosine coefficient* ini lebih unggul diterapkan untuk proses peringkasan artikel berita pendek dibandingkan dengan metode *jaccard* dan *dice coefficient*. Dengan demikian, terbukti bahwa peringkasan artikel berita pendek lebih tepat menggunakan metode MMR berbasis *cosine coefficient.*

2. Saran

Peringkasan menggunakan metode *cosine coefficient* lebih baik dibandingkan dengan metode *boolean model* lainnya dan metode *vector space model* (TF-IDF-DF). Pada penelitian selanjutnya, peneliti memberikan saran sebagai berikut:

a. Pada penelitian ini, *text preprocessing* masih menggunakan perhitungan manual, sehingga membutuhkan waktu lebih lama untuk memprosesnya. Oleh karena itu, pada penelitian selanjutnya bisa dirancang sistem (program) *text preprocessing* khusus untuk teks berbahasa Indonesia, supaya lebih efisien waktu.
b. Antara hasil *f-measure* metode *jaccard, dice*, dan *cosine,* metode *cosine coefficient* memang lebih unggul, akan tetapi ketiga hasil akurasi tersebut hanya sedikit perbedaannya. Oleh karena itu, pada penelitian selanjutnya bisa menambahkan metode lain yang dapat meningkatkan akurasi peringkasan.

DAFTAR PUSTAKA

[1] W. Yulita and F. S. Pribadi, "The Implementation of Maximum Marginal Relevance Method on Online National and Local News Portal," vol. 7, pp. 21–25, 2015.

[2] V. Thada and V. Jaglan, "Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved *Document*s Using Genetic Algorithm," *Int. J. Innov. Eng. Technol.*, vol. 2, no. 4, pp. 202–205, 2013.

[3] J. Ramos, J. Eden, and R. Edu, "Using TF-IDF to Determine Word Relevance in *Document* Queries," *Processing*, 2003.

[4] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of Jaccard Coefficient for Keywords Similarity," *Int. MultiConference Eng. Comput. Sci.*, vol. I, pp. 380–384, 2013.

[5] N. Anuar and A. B. Sultan, "Validate Conference Paper Using Dice Coefficient," vol. 3, no. 3, pp. 139–145, 2010.

[6] N. Agarwal, M. Rawat, and M. Vijay, "Comparative Analysis Of Jaccard Coefficient and Cosine Similarity for Web *Document* Similarity Measure," *Int. J. Adv. Res. Eng. Technol.*, vol. 2, no. 5, pp. 18–21, 2014.

[7] Sugiyamto, B. Surarso, dan A. Sugiharto, "Analisa performa metode cosine dan jacard pada pengujian kesamaan *document*," *J. Masy. Inform.*, vol. 5, pp. 1–8, 2014.

[8] S. Xie and Y. Liu, "Using corpus and knowledge-based similarity measure in Maximum Marginal Relevance for meeting summarization," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, no. 2, pp. 4985–4988, 2008.

[9] Luhn, H.P., "The Automatic Creation of Literature Abstracts,"*IBM JOURNAL,* 159-165, 1958.

[10] Carbonell, J.G. dan J. Goldstein, "The Use of MMR and Diversity-Based Reranking in *Document* Reranking and Summarization," *ACM SIGIR conference on Research and development in information retrieval*, 12:335-336, 1998.

[11] Mulyana, I., S. Ramadhona, dan Herfina, "Penerapan Terms Frequency-Inverse *Document* Frequency Sistem Peringkasan Teks Otomatis *Document* Tunggal Berbahasa Indonesia," *KNASTIK*, 1-8, 2012.

[12] Yusintan, B. P., Y. Firdaus, dan W. Maharani,"Perangkingan Ulang *Document* Teks dengan Metode Maximal Marginal Relevance untuk Menghasilkan Ringkasan Teks dengan Redundansi Minimum,"*Tel-U Collection,* 1-2, 2010.

[13] Nazief, B. A. A. and M. Adriani,"Confix-Stripping : Approach to Stemming Algorithm for Bahasa Indonesia,"*International Conference on Informationand Knowledge Management,* 560-563, 1996.

[14] Pramono, L.H., A.S. Rohman, and H. Hindersah,"Modified WeightingMethod in TF*IDF Algorithm for Extracting User Topic Based on Email and Social Media in Integrated Digital Assistant,"*Rural Information & Communication Technology and Electric-Vehicle Technology,* 1-6, 2013.

[15] Nedunchelian, R., R. Muthucumarasamy, and E. Saranathan, "Comparison of Multi *Document* Summarization Techniques," *International Journal of Computer Applications*, 11(3) : 155-160, 2011.

[16] Aditya, CSK.,"Vector Space Model (VSM) dan Pengukuran Jarak pada Information Retrieval (IR)", [Online] Available at: https://informatikalogi.com/vector-space-model-pengukuran-jarak/ 2016, [Accessed 22 Juni 2017].

[17] Romli, Asep Syamsul M.**,** "Teknik Menulis di Media Online - Jurnalistik Online", [Online] Available at: http://www.romelteamedia.com/2014/06/teknik-menulis-di-media-online.html, [Accessed 22 Juni 2017].