Seleksi Fitur Bobot Kata dengan Metode TFIDF untuk Ringkasan Bahasa Indonesia

Ni Komang Widyasanti, I Ketut Gede Darma Putra, Ni Kadek Dwi Rusjayanthi
Program Studi Teknologi Informasi, Fakultas Teknik, Universitas Udayana
Bukit Jimbaran, Bali, Indonesia Telp. (0361) 701806
e-mail: widyas96@gmail.com, ikgdarmaputra@unud.ac.id , dwi.rusjayanthi@unud.ac.id

Abstrak

Penyebaran informasi dalam bentuk teks digital semakin tak terbendung seiring perkembangan waktu. Kebutuhan akan membaca informasi juga tidak pernah berkurang, berdasarkan riset yang dilakukan pada lima kota besar di Indonesia sepanjang tahun 2015 oleh okezone.com menyatakan persentasi konsumsi berita secara online mencapai 96%. Salah satu solusi untuk mempermudah dan mempercepat pencarian informasi yang sesuai adalah dengan meringkas konten tersebut. TFIDF (Term Frequency Inverse Document Frequency) merupakan metode pembobotan dalam bentuk integrasi antar term frequency dengan inverse document frequency. Metode TFIDF digunakan pada penelitian ini untuk memilih fitur sebagai hasil ringkasan, dengan penerapannya pada seleksi fitur bobot kata. Nilai kepuasan pembaca sebesar 61,94%. Durasi ringkasan rata-rata 68,25 detik dengan jumlah kalimat dan kata rata-rata 31,875 dan 387,375. Penelitian dilakukan menggunakan jenis dokumen fiksi dan non-fiksi serta seleksi fitur disetiap paragrafnya, yang membedakannya dengan penelitian terkait sebelumnya.

Kata Kunci: *Ringkasan Teks Otomatis, Pembobotan TFIDF, Bahasa Indonesia*

Abstract

The dissemination of information in the form of digital text is growing rapidly over the time. The need to read the information also never diminished, based on research conducted on five major cities in Indonesia during 2015 by okezone.com states the percentage of online news consumption reached 96%. One solution to simplify and speed up the search for appropriate information is to summarize the content. TFIDF (Term Frequency Inverse Document Frequency) is a method of weighting in the form of integration between term frequency with inverse document frequency. The TFIDF method is used in this research to select the feature as a summary result, with its application on feature selection of term weight. Reader satisfaction score of 61.94%. The average summary computation is 68.25 seconds with the average number of sentences and words of 31.875 and 387.375*. This research conducted with fiction and non-fiction documents, also with its feature selection for every paragraph that makes it different from the previous research.*

Keywords*:* *Automatic Text Summarization, TFIDF Weighting, Indonesian Language*

1. Pendahuluan

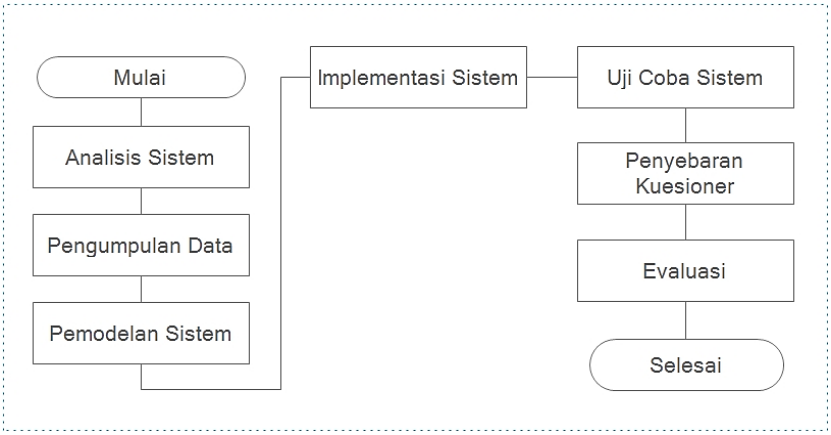
Penyebaran informasi di dunia maya semakin berkembang pesat, seiring dengan perkembangan teknologi. Riset yang dilakukan oleh *okezone.com* di lima kota besar di Indonesia sepanjang tahun 2015 meyatakan bahwa konsumsi berita melalui online mencapai 96%. Angka tersebut merupakan persentase tertinggi, dibandingkan dengan konsumsi berita melalui televisi (91%), surat kabar (31%), dan radio (15%) [1]. Informasi yang terdiri dari berbagai macam kategori dan tersebar acak begitu saja terkadang menyulitkan masyarakat yang membutuhkan informasi, belum lagi panjangnya informasi tidak seutuhnya berisi hal penting yang dibutuhkan pembaca. Solusi dari permasalah tersebut adalah dengan meringkas teks. Ringkasan teks pada penelitian ini dilakukan dengan menggunakan metode TFIDF *(Term Frequency Inverse Document Frequency)* yang merupakan salah satu bagian dari pendekatan *Frequency-Driven* [2]. Metode TFIDF diaplikasikan kedalam pembobotan berbasis kata sebagai fitur hasil ringkasan.

Penelitian sebelumnya yang membahas mengenai ringkasan teks diantaranya adalah "*Text Summarization* untuk Dokumen Berita Berbahasa Indonesia" oleh Romadhony dan kawan-kawan yang membahas mengenai *text summarization* dengan menggunakan metode *extraction* serta *stemming* yang menggunakan Algoritma Nazief dan Adriani yang telah dimodifikasi. Hasil menunjukkan bahwa akurasi tertinggi didapatkan pada *compression rate* 50% sementara durasi komputasi tercepat pada *compression rate* 25%, hanya saja hasil yang didapat masih belum maksimal dan *compression rate* tidak mempengaruhi durasi komputasi [3]. "*Frequent Term based Text Summarization for Bahasa Indonesia*" oleh Fachrurrozi dan kawan-kawan merupakan sebuah penelitian terkait *text summarization* untuk Bahasa Indonesia berbasi *frequent term* yang diimplementasikan di Java. Sistem menghasilkan ringkasan berdasarkan identifikasi dan ekstraksi kalimat penting pada dokumen *input*. Hasil uji coba menyatakan hasil ringkasan memiliki nilai F-measure sebesar 78% pada rasio kompresi 30%, dan rata-rata nilai dari responden sebesar 83,3 [4]. "*Comparing Fuzzy Logic and Fuzzy C-Means (FCM) on summarizing indonesian language document*" oleh Riandayani dan kawan-kawan menghasilkan ringkasan dengan akurasi terendah sebesar 50,83% dan 53,33% dengan Fuzzy C-Means dan Fuzzy Logic secara berurutan pada rasio kompresi 20% serta akurasi tertinggi sebesar 63,75% dan 64,17% pada rasio kompresi 35%. Durasi kompresi rata-rata 56,4 dan 42,7 detik untuk metode Fuzzy Logic dan Fuzzy C-Means pada jumlah 260 kata [5].

Penelitian terkait peringkasan teks otomatis yang dikembangkan dilakukan dengan menggunakan 8 dokumen uji yang terdiri dari dokumen fiksi (cerita rakyat dan dongeng) serta *non*-fiksi (berita dan artikel) dengan rata-rata jumlah kata sebanyak 387,375 dan kalimat sebanyak 31,875. Ringkasan yang dihasilkan dievaluasi berdasarkan *compression* rate, kepuasan pembaca serta durasi komputasi. Fokus penelitian ini terletak pada hasil penerapan metode TFIDF dalam basis bobot kata dengan seleksi fitur terhadap setiap paragrafnya pada jenis dokumen fiksi dan non fiksi, yang membedakannya dari penelitian terkait sebelumnya.

2. Metodologi Penelitian

Penelitian terkait penerapan metode TFIDF pada ringkasan otomatis dokumen bahasa indonesia berbasis bobot kata dilakukan dalam tujuh langkah yang dapat dilihat pada Gambar 1.



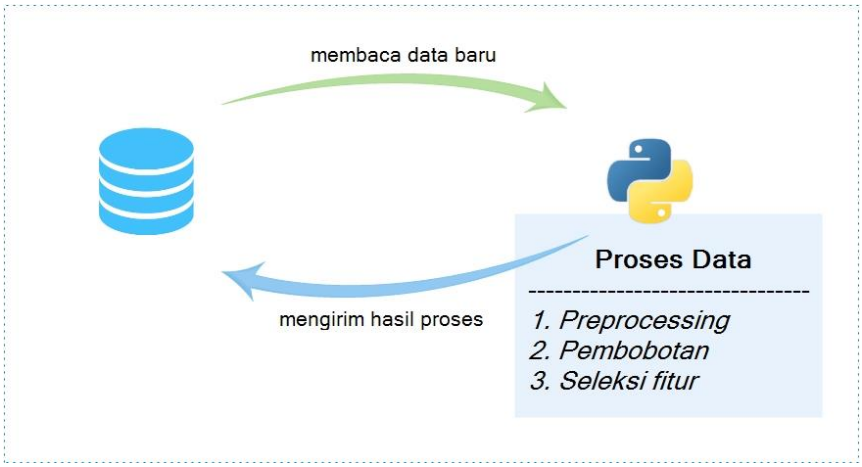
Gambar 1. Alur Penelitian

Tujuh langkah penelitian pada Gambar 1 dimulai dengan menganalisis sistem, yaitu tahap melakukan identifikasi dan analisis yang lebih spesifik terkait peringkasan teks otomatis dan penerapan metode TFIDF berbasis bobot kata. Langkah kedua yakni pengumpulan data, yaitu tahap pengumpulan dokumen dalam bentuk cerita rakyat, dongeng, berita, dan artikel yang digunakan sebagai data uji. Langkah ketiga adalah pemodelan sistem, yaitu pembuatan model data dan peringkasan teks menggunakan metode TFIDF yang diterapkan pada basis bobot kata. Langkah keempat mengimplementasikan sistem berbasis *web* dengan mempresentasikan hasil desain ke dalam aplikasi DBMS dengan bantuan pemrosesan menggunakan Bahasa Pemrograman Python. Langkah kelima adalah proses uji coba sistem untuk mengetahui tingkat keberhasilan dan kehandalan sistem yang dibuat secara keseluruhan.

Langkah keenam yakni melakukan penyebaran kuesioner untuk mengetahui tingkat kepuasan pembaca terhadap hasil peringkasan dokumen otomatis oleh sistem. Langkah terakhir adalah mengevaluasi hasil penelitian dari segi durasi komputasi, kepuasan pembaca, serta menarik simpulan dan memberi saran atas kendala yang ditemukan dalam pembuatan sistem.

2.1 Gambaran Umum Sistem

Penelitian terkait ringkasan otomatis teks Bahasa Indonesia ini secara umum dapat digambarkan seperti pada Gambar 2.

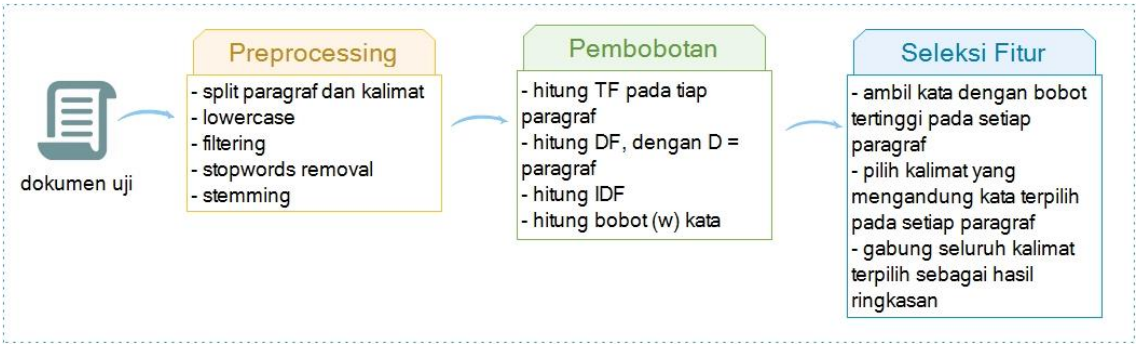


Gambar 2. Gambaran Umum Sistem

Gambar 2 menjelaskan sistem yang dirancang menggunakan *engine* Python untuk memproses dokumen, serta MySQL sebagai DBMS. *Engine* Python mengecek apakah terdapat data baru pada *database* yang belum diolah sebelum menjalankan proses berupa *preprocessing,* pembobotan, dan seleksi fitur untuk dijadikan ringkasan. Data yang telah diproses menjadi ringkasan kemudian dilempar kembali ke *database*.

2.2 Seleksi Fitur

Ringkasan berbasis bobot kata didapatkan berdasarkan nilai bobot kata tertinggi dari masing-masing paragraf yang terdapat pada dokumen. Proses lebih rinci dapat dilihat pada Gambar 3.



Gambar 3. Proses pada Ringkasan Berbasis Bobot Kata

Gambar 3 merupakan rincian proses yang dilakukan untuk mendapat hasil ringkasan berbasis bobot kata, dimana proses awal yang terjadi adalah *preprocessing* dokumen dengan langkah pertama membagi dokumen menjadi paragraf dan kalimat, kemudian mengubah teks menjadi *lowercase.* Langkah berikutnya adalah melakukan *filter* terhadap karakter *non*-huruf, yang kemudian dilanjutkan dengan menghapus kata yang termasuk pada *stop words*. Langkah terakhir adalah melakukan *stemming* dengan algoritma Sastrawi. Dokumen yang telah melalui *preprocessing* kemudian dipecah menjadi kata untuk dilakukan pembobotan TFIDF atau proses *term weighting,* dengan catatan *D* yang seharusnya dokumen diadaptasi menjadi paragraf. Kata

yang memiliki bobot tertinggi di masing-masing paragraf dipilih, untuk pemilihan kalimat yang mengandung kata tersebut. Kalimat-kalimat yang terpilih digabungkan menjadi satu sebagai hasil ringkasan.

3 Kajian Pustaka

Kajian pustaka memuat materi yang menjadi referensi penelitian ini. Referensi yang dimuat yakni terkait *text mining, text summarization, text preprocessing,* dan TFIDF.

3.1 Text Mining

Text mining didefinisikan sebagai sebuah proses menggali informasi, dimana pengguna berinteraksi dengan dokumen-dokumen menggunakan alat analisis yang berupa komponen *data mining* yang diantaranya adalah komponen kategorisasi. *Text mining* bisa memberikan solusi atas berbagai masalah seperti *preprocessing*, pengelompokan, hingga analisa teks yang tidak terstruktur dalam jumlah yang besar. *Text mining* mengadopsi berbagai teknik dari bidang lain, seperti *Data Mining, Information Retrieval*, *Machine Learning,* statistik dan matematik, *linguistic, Natural Languange Processing* (NLP), serta *visualization*. Kegiatan terkait riset untuk *text mining* diantaranya adalah ekstraksi dan penyimpanan *text*, *preprocessing*, pengumpulan data statistik, *indexing,* dan analisis konten [6]. *Text mining* digunakan untuk mengolah data yang bersifat tidak terstruktur (seperti artikel, teks dari web, *blog post,* dan lain sebagainya), berbeda dengan *data mining* yang lebih cenderung digunakan untuk mengolah data yang terstruktur. *Data mining* merupakan langkah yang terdiri dari penerapan analisis data dan penemuan algoritma yang menghasilkan enumerasi tertentu pada pola data [7].

3.2 Text Summarization

Summarization atau ringkasan merupakan teks yang dihasilkan dari sebuah teks atau kumpulan teks yang mengandung isi berupa informasi dari teks asli dengan panjang yang tidak lebih dari setengah panjang teks aslinya [8]. Penelitian peringkasan teks otomatis dibentuk oleh Luhn sejak tahun 1958. Teknik-teknik yang digunakan dalam peringkasan diantaranya ialah teknik pendekatan statistika seperti teknik frekuensi kemunculan kata oleh Luhn tahun 1958, teknik posisi kata oleh Baxendale tahun 1958, kata kunci dan *heading* oleh Edmudson tahun 1969, posisi kalimat oleh Lin dan Hoovy tahun 1997, dan teknik pendekatan dengan analisis bahasa natural seperti *inverse term frequency* (itf) and teknik *NLP* oleh Aone tahun 1990, *lexical chain* oleh Mc Keown tahun 1997, dan relevansi maksimal marginal oleh Cabonell dan Goldstein tahun 1998.

3.3 Text Preprocessing

Tujuan dari pemrosesan awal atau *preprocessing* adalah untuk mempersiapkan text menjadi data yang siap diproses. Proses yang dilakukan pada tahap ini meliputi *case folding* atau mengubah seluruh huruf yang ada pada dokumen menjadi huruf kecil serta menghilangkan karakter selain huruf, *filtering* atau tahap mengambil kata penting dengan cara menghilangkan *stopwords*, *tokenization* atau tahapan dimana kumpulan karakter dalam teks yang telah melalui proses *case folding* akan dipecah kedalam satuan kata (*token*), dan *stemming* yang merupakan untuk mentransformasikan kata-kata yang terdapat dalam dokumen ke kata-kata akar atau dasarnya dengan menggunakan berbagai aturan tertentu [9], [10].

3.4 TFIDF

TFIDF adalah sebuah metode yang merupakan integrasi antar *term frequency* (TF), dan *inverse document frequency* (IDF). *Term Frequency* dihitung menggunakan Persamaan (2) dengan *term frequency* ke-i adalah frekuensi kemunculan term ke-i dalam dokumen ke-j. *Inverse Document Frequency* (IDF) adalah logaritma dari rasio jumlah seluruh dokumen dalam korpus dengan jumlah dokumen yang memiliki term yang dimaksud seperti yang dituliskan secara matematis pada Persamaan (3). Nilai didapatkan dengan mengalikan keduanya yang diformulasikan pada Persamaan (4).

$$tf_i = \frac{freq_i(d_j)}{\sum_{i=1}^{k} freq_i(d_j)} \quad (2)$$

$$idf_i = log \frac{|D|}{|\{d:t_i \in d\}|} \quad (3)$$

$$(tf - idf)_{ij} = tf_i(d_j) * idf_i \quad (4)$$

Fungsi metode TFIDF adalah untuk mencari representasi nilai dari tiap-tiap dokumen dari suatu kumpulan data *training* (*training set*) dimana nantinya dibentuk suatu vektor Antara dokumen dengan kata (documents with *terms*) yang kemudian untuk kesamaan antar dokumen dengan *cluster* akan ditentukan oleh sebuah *prototype* vektor yang disebut juga dengan *cluster centroid* [11], [12], [13].

4. Hasil dan Pembahasan

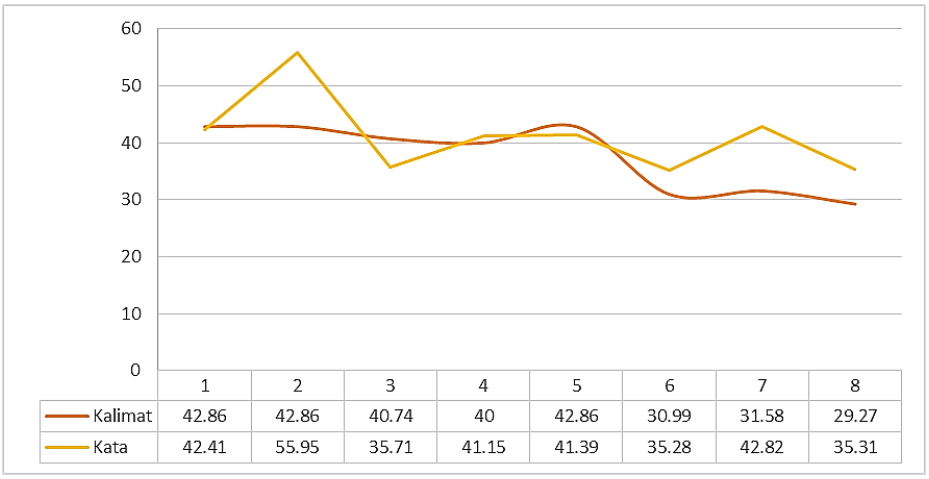
Penelitian dilakukan dengan menggunakan 8 dokumen uji yang terdiri dari 4 teks fiksi dan 4 teks non-fiksi, dengan rincian seperti pada Tabel 1.

Tabel 1. Rincian Dokumen Uji

Banyak Dokumen	8
Rata-rata kalimat per-dokumen	31,875
Jumlah maksimum kalimat per-dokumen	71
Jumlah minimum kalimat per-dokumen	14
Jumlah maksimum kata per-dokumen	703
Jumlah minimum kata per-dokumen	168

4.1 Evaluasi *Compression Rate*

Hasil *compression rate* dari ringkasan berbasis bobot kata pada seluruh dokumen uji dapat dilihat secara detail pada Gambar 4.

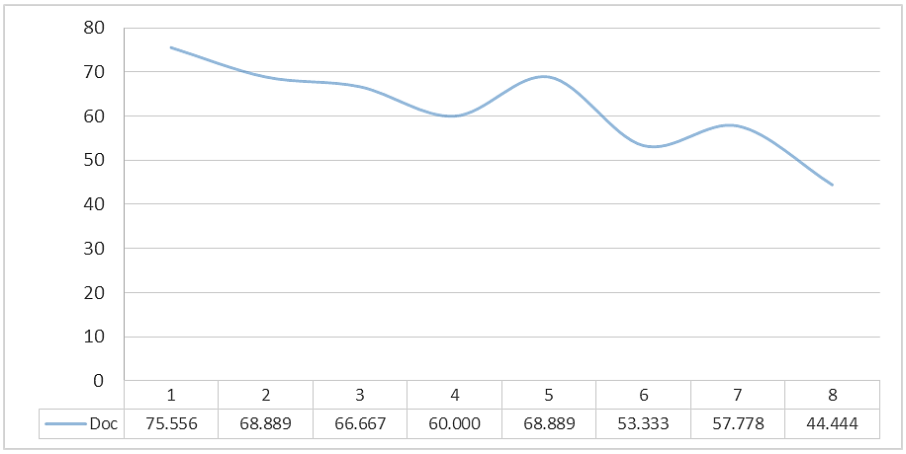


Gambar 4. Grafik Compression Rate

Gambar 4 merupakan grafik rincian hasil *compression rate* kalimat dan kata dari masing-masing dokumen yang diuji, dengan nilai kompresi terendah untuk kalimat adalah 29,27% dan tertinggi adalah 42,86%. Nilai kompresi terendah untuk kata adalah 35,28% dan tertinggi adalah 55,95%.

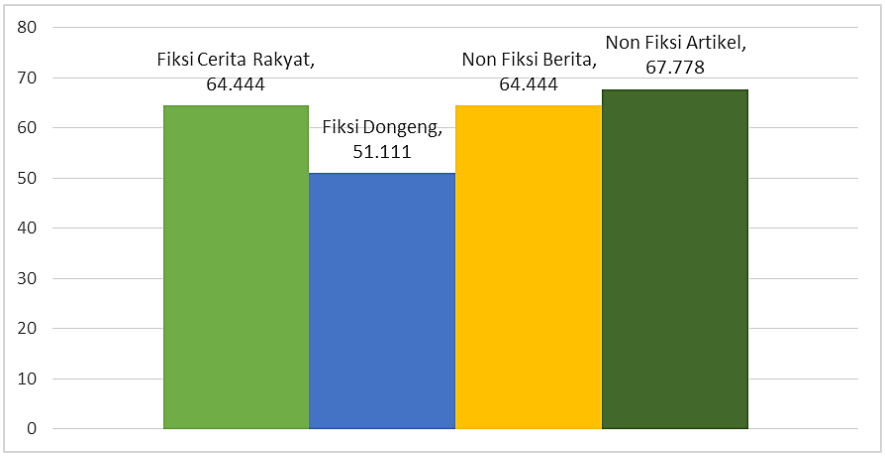
4.2 Evaluasi Kepuasan Pembaca

Nilai kepuasan pembaca didapat dengan cara penyebaran kuesioner kepada 72 responden. Pembaca diberi kuesioner dengan rentang nilai 1-5. Hasil rata-rata kuesioner kepuasan pembaca dapat dilihat pada Gambar 5.



Gambar 5. Grafik Nilai Kepuasan Pembaca

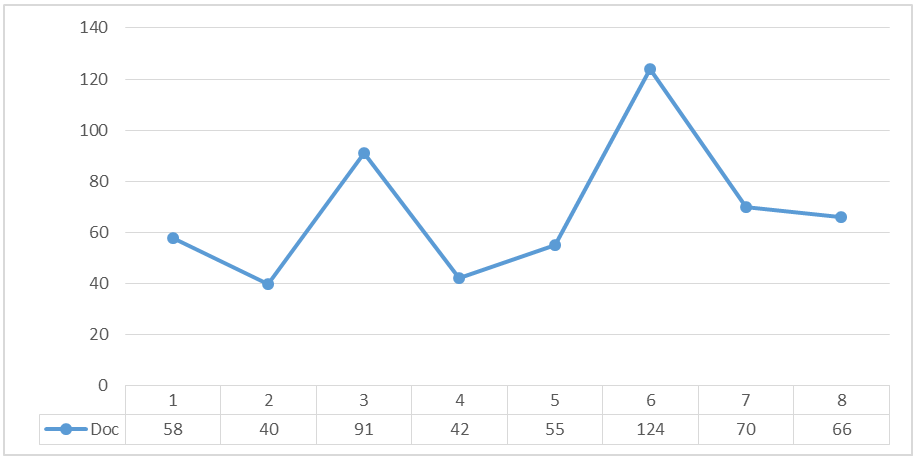
Rata-rata nilai kepuasan pembaca seperti pada Gambar 5 menunjukkan nilai tertinggi pada ringkasan adalah 75,556% dan terendah adalah 44,444%. Hasil rata-rata kuesioner kepuasan pembaca berdasarkan jenis dokumennya dapat dilihat pada Gambar 6. Nilai tertinggi rata-rata nilai kepuasan pembaca adalah dokumen jenis non-fiksi artikel, dengan nilai 67,78% dan terendah adalah dokumen jenis fiksi dongeng dengan nilai 51,11%.



Gambar 6. Grafik Rata-rata Nilai Kepuasan Pembaca

4.4 Evaluasi Durasi Komputasi

Durasi komputasi dari ringkasan berbasis bobot kata dan kalimat pada seluruh dokumen uji dapat dilihat secara detail pada Gambar 7.



Gambar 7. Grafik Durasi Komputasi

Gambar 7 merupakan grafik rincian hasil durasi komputasi, waktu yang dihasilkan berdasarkan dengan banyaknya kata dan kalimat yang dikandung oleh teks. Durasi tercepat adalah selama 40 detik, untuk dokumen yang mengandung 14 kalimat dan 168 kata, sementara durasi terlama adalah selama 124 detik untuk dokumen yang mengandung 71 kalimat dan 703 kata.

5. Kesimpulan

Simpulan yang dapat diambil terkait penerapan metode TFIDF pada ringkasan otomatis dokumen Bahasa Indonesia untuk jenis fiksi dan *non*-fiksi berbasis bobot kata terhadap 8 dokumen uji dapat dibagi dalam empat bagian, berdasarkan *compression rate*, kepuasan pembaca, dan durasi komputasi. Simpulan evaluasi secara detail dapat dilihat pada Tabel 2.

Tabel 2. Simpulan Uji Coba

No	Keterangan	Fiksi	Non-Fiksi	Keseluruhan	satuan
1	Kalimat	44,5	19,25	31,875	kalimat
2	Kata	474,25	300,5	387,375	kata
3	*Compression Rate* Kalimat	33,68	41,62	37,645	%
4	*Compression Rate* Kata	38,96	43,55	41,2525	%
5	Durasi	79,5	57	68,25	detik
6	Kuesioner	57,78	66,11	61,94	%

Tabel 2 merupakan simpulan yang dapat diambil dari penelitian, dimana secara keseluruhan uji coba mengandung rata-rata 31,875 kalimat dan 387,375 kata. *Compression rate* rata-rata untuk kalimat adalah 37,645% dan 41,2525% untuk kata. Berdasarkan uji coba, ringkasan cenderung menghasilkan hasil akhir lebih banyak untuk dokumen *non*-fiksi. Durasi rata-rata untuk seluruh dokumen uji yakni 68,25 detik. Rata-rata kuesioner memiliki nilai 61,94%, dengan nilai dokumen *non*-fiksi kembali lebih tinggi dibanding dengan dokumen fiksi. Berdasarkan seluruh hasil uji coba yang dipaparkan, dapat dikatakan bahwa hasil ringkasan dengan metode TFIDF pada seleksi fitur bobot kata cukup baik, dan lebih cocok digunakan pada dokumen *non*-fiksi.

Daftar Pustaka

[1] Dedy Afrianto, “96% Masyarakat Indonesia Konsumsi Berita Online,” 2016.
[2] M. Allahyari *et al.*, “Text Summarization Techniques: A Brief Survey,” *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, pp. 398–405, 2017.
[3] A. Romadhony, Z. R. Fariska, N. Yusliani, and L. Abednego, “Text Summarization untuk

Dokumen Berita Berbahasa Indonesia," *Journal of Telkom University*, pp. 408–414, 2017.
[4] M. Fachrurrozi, N. Yusliani, and R. U. Yoanita, "Frequent Term based Text Summarization for Bahasa Indonesia," *International Conference on Innovations in Engineering and Technology*, pp. 30–32, 2013.
[5] D. A. Riandayani, I. K. G. Darma Putra, and P. W. Buana, "Comparing Fuzzy Logic and Fuzzy C-Means (FCM) on summarizing indonesian language document," *Journal of Theoretical and Applied Information Technology*, vol. 59, no. 3, pp. 718–724, 2014.
[6] N. M. A. Lestari and M. Sudarma, "Perencanaan Search Engine E-commerce dengan Metode Latent Semantic Indexing Berbasis Multiplatform," *Lontar Komputer*, vol. 8, no. 1, pp. 31–40, 2017.
[7] A. S. Devi, I. K. G. Darma Putra, and I. M. Sukarsa, "Implementasi Metode Clustering DBSCAN pada Proses Pengambilan Keputusan," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 6, no. 3, p. 185, 2015.
[8] C. C. Aggarwal and C. X. Zhai, *Mining text data*, vol. 9781461432. 2013.
[9] H. R. Pramudita, "Penerapan Algoritma Stemming Nazief & Andriani dan Similarity Pada Penerimaan Judul Thesis," *Jurnal Ilmiah DASI*, vol. 15, no. 4, pp. 15–19, 2014.
[10] N. M. A. Lestari, I. K. G. Darma Putra, and A. A. K. Cahyawan, "Personality Types Classification for Indonesian Text in Partners Searching Website Using Naïve Bayes Methods," *International Journal of Computer Science (IJCSI)*, vol. 10, no. 1, pp. 1–8, 2013.
[11] P. M. Prihatini, I. K. G. Darma Putra, I. A. Dwi Giriantari, and M. Sudarma, "Fuzzy-Gibbs Latent Dirichlet Allocation Model for Feature Extraction on Indonesian Documents," *Contemporary Engineering Sciences (CES)*, vol. 10, no. 9, pp. 403–421, 2017.
[12] M. N. Saadah, R. W. Atmagi, D. S. Rahayu, and A. Z. Arifin, "Sistem Temu Kembali Dokumen Teks dengan Pembobotan Tf-Idf Dan LCS," *Jurnal Ilmiah Teknologi Informasi (JUTI)*, vol. 11, no. 1, pp. 17–20, 2013.
[13] I. N. S. Paliwahet, I. M. Sukarsa, and I. K. G. Darma Putra, "Pencarian Informasi Wisata Daerah Bali menggunakan Teknologi Chatbot," *Lontar Komputer*, vol. 8, no. 3, pp. 144–153, 2017.