

LATENT SEMANTIC ANALYSIS (LSA) DAN AUTOMATIC TEXT SUMMARIZATION (ATS) DALAM OPTIMASI PENCARIAN ARTIKEL COVID 19

Herlina Jayadianti^{(1)*}, Ruth Damayanti⁽²⁾, Juwairiah⁽³⁾

^(1,2,3)Jurusan Teknik Informatika, Universitas Pembangunan Nasional Veteran Yogyakarta
Jl. Babarsari No.2, Janti, Caturtunggal, Kec. Depok, Kabupaten Sleman, Daerah Istimewa
Yogyakarta 55281

e-mail : herlinajayadianti@gmail.com, ruthdamayanti28@gmail.com , juwai_riah@gmail.com

Abstrak

The Covid 19 pandemic has given much awareness to all people around the world about the importance of maintaining health and changing lifestyles and lifestyles to be healthier. Clear, correct and precise information is indispensable to provide insight into this respiratory virus. Digital media is widely used by the public to find links about the covid19 virus. Health topics about covid 19 from several sites will be collected by scrapping method, and the data retrieval results will be processed to become an automatic summary using Latent Semantic Analysis (LSA), where this method, will help to find the hidden meaning of a collection of sentences. The formation of the summary is assisted by the cross method. The system also has an article search to allow users to find the right information. The results of this study showed that LSA method assisted by the cross method could be used in automatic summary shrinking well, test results in f-measure and recall values on average of 90.68% and 85% with the percentage of trained data: test data is 90:10. Data collection conducted during February-June 2020 was taken 120 training documents, and 12 test documents. Testing is done with a compression rate of 30%

Keywords: *automatic summary, health article, scrapping, latent semantic analysis, singular value decomposition, cross method*

Pandemi Covid 19 telah memberikan banyak kesadaran pada seluruh masyarakat dunia mengenai pentingnya menjaga kesehatan dan merubah pola hidup dan gaya hidup menjadi lebih sehat. Informasi yang jelas, benar dan tepat sangat diperlukan untuk memberi wawasan tentang virus pernafasan ini. Media digital banyak dipakai oleh masyarakat untuk mencari tautan mengenai virus covid19. Topik kesehatan mengenai covid 19 dari beberapa situs akan dikumpulkan dengan metode scrapping, dan hasil pengambilan data akan diolah untuk menjadi sebuah ringkasan otomatis dengan menggunakan Latent Semantic Analysis(LSA), dimana metode ini, akan membantu untuk menemukan makna tersembunyi dari sebuah kumpulan kalimat.Pembentukan ringkasan dibantu dengan metode cross method. Sistem ini juga memiliki sebuah pencarian artikel, untuk membuat pengguna dapat menemukan informasi secara tepat. Hasil dari penelitian ini menunjukkan bahwa metode LSA yang dibantu dengan cross method dapat digunakan dalam penyusunan ringkasan otomatis dengan baik, Hasil pengujian menghasilkan nilai f-measure dan recall rata-rata sebesar 90.68% dan 85% dengan presentase data latih: data uji adalah 90:10. Pengumpulan data dilakukan selama bulan Februari-Juni 2020 diambil 120 dokumen latih, dan 12 dokumen uji. Pengujian dilakukan dengan compression rate sebesar 30%

Kata kunci: *ringkasan otomatis, berita kesehatan, scrapping, Latent Semantic Analysis,Singular Value Decomposition, Cross Method.*

1. PENDAHULUAN.

Dunia sedang mengalami krisis yang belum pernah terjadi sebelumnya karena covid-19 breakout, dan masyarakat di seluruh dunia terpaksa tinggal di dalam rumah untuk keselamatan. Dalam situasi seperti itu, meningkatnya dan turunnya jumlah kasus atau kematian yang terkena dampak telah berubah menjadi berita utama yang konstan di sebagian besar saluran berita. Akibatnya, orang kebanjiran berita Covid 19 (Chakraborty,2020). Pemberitaan mengenai Covid19 sangat banyak dan beragam dan ini telah berlangsung terus menerus sejak Desember 2019. Banyaknya jumlah platform online yang memuat tentang berita tentang Covid 19, membuat pembaca berita sulit untuk menemukan informasi yang tepat dengan waktu yang singkat mengenai hal seputar virus ini. Pengguna tidak mungkin untuk mengunjungi satu per satu situs untuk mendapatkan informasi yang diinginkan, karena saat ini seluruh situs berita menampilkan berita yang sama mengenai wabah pandemic (L Singh, S Bansal, 2020) Pemberitaan itu justru menakutkan namun disatu sisi pembaca ingin mencari karena rasa ingin tahunya. Pemberitaan di media online yang tiada henti, banyak memberi edukasi bagaimana menghadapi virus itu sejak awal, namun sayangnya butuh waktu cukup banyak untuk membaca seluruh berita tersebut. Pembaca tentu sangat menginginkan adanya ringkasan bagaimana berita Covid19 dikumpulkan secara otomatis, diringkas dan disajikan.

Fokus penelitian yang akan dikerjakan adalah bagaimana meringkas berita terkait covid19 dengan cepat dan tepat. Tahapan yang akan dilakukan adalah mengumpulkan semua data mengenai covid19 dari situs-situs yang telah ditentukan sebagai objek penelitian sumber berita yakni Kompas.com, Liputan6.com, Klikdokter.com, detikhealth.com. Penelitian ini akan melakukan peringkasan teks otomatis dengan memperhatikan penggunaan semantik. Metode Latent Semantic Analysis (LSA) akan digunakan untuk menyelesaikan masalah ekstraksi kata kunci. Metode ini dapat memberikan ringkasan otomatis dan juga mengekstraksi kata kunci yang dapat membantu pengoptimalan pencarian. Setiap situs berita Kompas.com, Liputan6.com, Klikdokter.com, dan detikhealth.com menggunakan penyusunan dan pemilihan kata / terminology yang berbeda sedangkan pengetahuan pengguna dalam mencari berita sangat terbatas. (M Cinelli,, 2020).

Pencarian berita dan peringkasan teks otomatis dengan metode *maximum marginal relevance* dilakukan oleh Mustaqhfi. *Maximum marginal relevance* digunakan dalam peringkasan teks otomatis tanpa disertai dengan sebuah mesin pencari (Mustaqhfi et al., 2012), sehingga terdapat suatu kendala yaitu pengguna harus satu persatu memilih file yang akan diringkas. Contoh lain yaitu penelitian tentang sistem temu kembali dengan metode *vector space model* (Irmawati, 2017) penelitian ini hanya memberikan sistem temu kembali tanpa menyediakan ringkasan teks otomatis, sehingga untuk mendapatkan informasi pengguna harus menelusuri situs satu persatu. Dibandingkan dengan *maximum marginal relevance* dan *vector space model*, *Latent Semantic Analysis* (LSA) dianggap lebih baik karena dapat mereduksi kalimat pada artikel tanpa menghilangkan makna keseluruhan artikel. Hipotesa pada penelitian ini adalah LSA dapat menyediakan sistem temu kembali dan peringkasan teks otomatis secara bersamaan, sehingga mempermudah pengguna dalam mendapatkan informasi. *Scraping* merupakan suatu kegiatan untuk mengekstraksi data dari website yang diinginkan. Kemampuan ekstraksi data yang dimiliki, diterapkan pada sistem ini untuk pengambilan informasi dari suatu link artikel seperti judul artikel, penulis, tanggal terbit, dan isi artikel. Artikel yang telah diambil, nantinya akan diolah menjadi sebuah ringkasan teks guna mempermudah pengguna untuk menemukan intisari informasi. Metode yang digunakan dalam proses peringkasan teks adalah metode

2. METODOLOGI PENELITIAN

2.1. *Crawling*

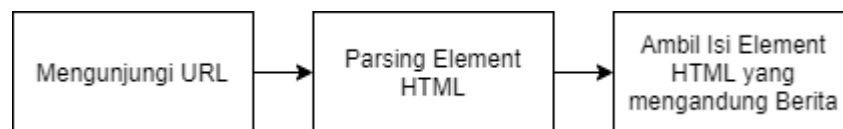
Proses *crawling* merupakan proses awal tahapan untuk *information extraction* yang akan diimplementasikan pada *website* terkait, dimana hasil dari ekstraksi informasi akan diolah dan dianalisis secara lanjut sesuai dengan kebutuhan, yaitu membuat ringkasan. Proses *crawling* dimulai dengan list URL yang akan dikunjungi, disebut *seeds*, dimana *seeds* akan ditentukan secara mutlak oleh pengguna untuk memulai proses *crawling*, List URL yang digunakan sudah merupakan URL absolute, sehingga URL yang digunakan bisa disimpan sebagai *frontier*, dan dapat langsung digunakan untuk *fetching page* URL, dalam proses ini jika ditemukan *link* dengan alamat halaman utama yang sama, maka akan secara otomatis ditambahkan sebagai *frontier* baru, dan akan dilakukan juga proses *fetching*, proses ini dilakukan secara rekursif dengan beberapa

kebijakan tertentu (Shkapenyuk & Suel, 2002) . *Frontier* pada penelitian ini berisi tentang informasi artikel memuat judul, penulis, tanggal penerbitan.

Proses untuk melakukan *information extraction* dari *frontier* terkait menggunakan metode *scrapping*. Proses ini bisa digambarkan seperti kegiatan mengcopy dan *paste* informasi, namun dilakukan secara otomatis. Teknik untuk melakukan scraping ada beberapa cara dengan cara HTTP programming, DOM parsing, dan HTML parser (Vargiu & Urru, 2012). Pada penelitian ini digunakan dengan proses HTML parser, dimana sistem akan mendeteksi setiap *element* HTML yang ada, dan akan mengambil *element* yang mengandung berita didalamnya. Sebagai contoh, data yang akan diambil dari Kompas.com dapat dilihat pada tabel 1. Setiap *website* akan memiliki struktur tag HTML yang berbeda, sehingga saat *fetching page*, diperlukan untuk filtering tag HTML dan melakukan penyaringan untuk tag HTML yang dibutuhkan saja.

Tabel 1 Jenis Informasi dan Tag HTML

No	Informasi	Tag HTML
1	Judul Berita	read_title
2	Nama Penulis	read_credit_item_penulis
3	Tanggal Penerbitan	read_credit_item_date
4	Isi Berita	read_content



Gambar 1. Proses Scraping HTML Parser

2.2. Text Preprocessing

Tahap pra proses (*preprocessing*) adalah tahapan awal dalam peringkasan teks, di dalam tahapan ini ada beberapa tahapan turunan, yaitu pemecahan kalimat, *case folding*, *tokenizing*, penghapusan *stopwords*, dan *stemming*. Tahap ini, berfungsi untuk menyiapkan dokumen mentah menjadi dokumen yang siap diolah atau representatif. (Larocca Neto et al., 2000). Proses *stemming* pada penelitian ini menggunakan Sastrawi Stemmer, dimana Sastrawi mengimplementasikan Algoritma Nazief & Adriani, dimana algoritma ini adalah algoritma stemming untuk Bahasa Indonesia (Adriani et al., 2007).

2.3. TF-IDF

Proses TF-IDF merupakan proses yang penting dalam penelitian ini, dikarenakan hasil dari proses ini akan digunakan untuk matriks inputan yang akan dilakukan pada proses selanjutnya, proses ini akan menghitung bobot kata pada sebuah dokumen. Jika sebelumnya, pada proses *preprocessing* sebuah artikel telah dipecah menjadi potongan-potongan kalimat, dan potongan kalimat juga telah dipecah menjadi potongan *corpus*, sehingga pada proses ini, akan dicari bobot sebuah kata yang muncul dalam sebuah dokumen, untuk menemukan bobot tersebut, akan dihitung jumlah frekuensi kemunculan sebuah kata pada sebuah dokumen (*Term Frequency*) (Doko et al., 2013). Semakin besar kemunculan kata pada sebuah dokumen maka bobotnya akan semakin besar. Perhitungan bobot akan memperhitungkan kebalikan frekuensi dokumen yang mengandung sebuah kata (*Inverse Document Frequency*).

$$W = TF * (IDF + 1) \quad (1)$$

$$IDF = \log^D / df \quad (2)$$

Keterangan:

- TF : Frekuensi kemunculan kata pada setiap dokumen
D : Jumlah dokumen
df : Jumlah dokumen yang mengandung sebuah term
IDF : Hasil log dari jumlah dokumen keseluruhan dibagi dengan jumlah dokumen yang mengandung sebuah *term*
W : Bobot frekuensi sebuah term pada sebuah dokumen

Penelitian ini, menggunakan sebuah *library Scikit-learn*, dimana terdapat beberapa parameter yang bisa diatur, seperti *smooth_idf* dan *norm*. Parameter *smooth_idf* akan menambahkan satu buah konstanta dalam pembilang dan penyebut dari *IDF*, sehingga kata yang muncul di setiap dokumen tidak sepenuhnya diabaikan, dan menghindari pembagian nol. Penggunaan parameter *smooth_idf* dikombinasikan dengan penggunaan parameter *norm*, dimana hasil akan dinormalisasikan sesuai dengan *range* perhitungan manual, sehingga hasil perhitungan manual maupun dengan *library* sama.

2.4. Latent Semantic Analysis

Metode LSA dapat diterapkan pada sistem temu kembali, karena dapat melakukan pencarian dokumen yang relevan menurut sistem, namun relevansi merupakan hal yang subjektif. Pada umumnya dokumen akan dikatakan relevan dengan query apabila dokumen memuat kata yang sama atau dokumen mempunyai makna yang sama.

Contoh kasusnya adalah jika dokumen yang dianggap relevan yang memuat kata kunci saja maka jika terdapat **query 'corona'** maka dokumen yang akan dianggap relevan adalah dokumen yang memuat kata 'corona', dokumen yang memuat kata 'covid-19' akan dianggap tidak relevan. Sedangkan jika menilai relevansi dokumen berdasarkan makna maka dokumen yang mempunyai kata kunci 'covid-19' akan dianggap relevan karena mempunyai makna yang sama dengan kata kunci '**corona**'.

Metode LSA yang akan diimplementasikan pada sistem temu kembali akan mencari dan menemukan informasi sesuai dengan makna keseluruhan (*topic and conceptual meaning*) dari sebuah dokumen, sehingga bukan hanya berdasarkan kata, melainkan berdasarkan makna kata per kata, cara kerja LSA adalah dengan mengurangi *noise* yang terdapat pada sebuah kumpulan matriks dengan menggunakan SVD

LSA merupakan teknik statistika murni, bukan merupakan pemrosesan dari suatu *natural-language* atau kecerdasan buatan, tidak juga menggunakan susunan kamus, *knowledge-based*, *semantic-network* (jaringan semantik). Namun, metode ini hanya menggunakan inputan teks biasa yang diuraikan menjadi kata-kata lalu didefinisikan menjadi string karakter unik dan dipisahkan menjadi bagian bermakna yang berupa kalimat atau paragraph (Landauer et al., 1998).

2.5. Singular Value Decomposition

Metode SVD akan digunakan untuk mengurangi *noise* dari matriks masukan, dalam metode ini, matriks hasil yang akan diambil untuk proses selanjutnya adalah matriks VT dan S. Metode ini dapat mempengaruhi kecepatan komputasi dan peningkatan memori. (Martin & Porter, 2012)

$$A_{m \times n} = U_{m \times m} S_{m \times n} V_{n \times n}^T \quad (3)$$

Keterangan:

A : Matriks Masukan (Pada Penelitian matriks ini berisi matrik hasil perhitungan TF-IDF)

U&V : Matriks Ortogonal sehingga $U * U^T = I$ dan $V * V^T = I$, dimana I adalah matrik identitas

S : Matrik diagonal dengan elemen matriks positif atau nol

Proses akan digambarkan sebagai berikut:

Dokumen 1: Sebuah temuan terbaru didokumentasikan oleh tim peneliti dari Duke University belum lama ini.

Dokumen 2 :Mereka berhasil mengidentifikasi sejumlah mutasi diam pada sekitar 30.000 huruf kode genetik SARS-CoV-2.

Dokumen 3:Menurut peneliti, mutasi itu membuat virus corona menjadi pandemi global yang sulit untuk dihentikan.

Dokumen 4:Studi yang diterbitkan di Jurnal PeerJ, Jumat (08/05/2020), itu menyebutkan perubahan halus yang terjadi menunjukkan bagaimana virus dapat melipatgandakan molekul RNA dalam sel manusia.

Dokumen 5:Penulis utama penelitian itu, Alejandro Berrio, mengatakan perubahan struktur RNA itu kemungkinan dapat membedakan SARS-CoV-2 pada manusia dari virus corona lainnya.

Dokumen 6:Oleh karena itu, pandemi Covid-19 disebut lebih sulit dikendalikan dibandingkan wabah virus corona sebelumnya.

Gambar 2. Dokumen Masukan

Kumpulan dokumen akan dihitung dengan TF-IDF dan akan menjadi matriks masukan. Dengan menggunakan LSA yang diimplementasikan dengan SVD. Dalam SVD, matriks persegi panjang didekomposisi menjadi produk dari tiga matriks lainnya. Satu matriks komponen menggambarkan entitas baris asli sebagai vector turunan nilai faktor ortogonal, yang lain menjelaskan entitas kolom asli dalam hal yang sama cara, dan yang ketiga adalah matriks diagonal yang berisi nilai skala sedemikian rupa sehingga ketika tiga komponen dikalikan matriks, matriks asli direkonstruksi. Namun untuk membentuk suatu ringkasan, hanya dua buah komponen saja yang diperlukan yaitu komponen orthogonal yang menggambarkan entitas baris asli, dan komponen lainnya adalah matriks diagonal dengan nilai skala. (Landauer et al., 1998).

-0.10717539	0.10717539	-0.10513468	...	-0.15402925	-0.12222446	-0.10513468
0.03133021	0.03133021	-0.2185961	...	0.25514567	0.14434338	-0.2185961
-0.01381522	-0.01381522	0.22471936	...	0.01061278	0.06180603	0.22471936
-0.00718632	-0.01107943	-0.00895706	...	-0.08195549	-0.01342038	0.04952451
0.00084305	0.0013305	0.00104915	...	0.01588708	0.00129861	-0.00852281
-0.00646336	-0.00985835	-0.00806161	...	-0.04366308	-0.01302549	0.03514508

Gambar 3. Matriks VT

1.23879497	0	0	...	0	0	0
0	0.85319836	0	...	0	0	0
0	0	0.92208038	...	0	0	0
0	0	0	...	1.01022382	0	0
0	0	0	...	0	0.97929445	0
0	0	0	...	0	0	0.95270014

Gambar 4. Matriks S

2.6. Cross-Method

Cross Method adalah salah satu pengembangan metode dari ekstrasi kalimat setelah proses perhitungan SVD, dengan kata lain, untuk menentukan kalimat terpilih dari nilai matriks VT. Metode ini dikembangkan oleh (Steinberger & Ježek, 2009).

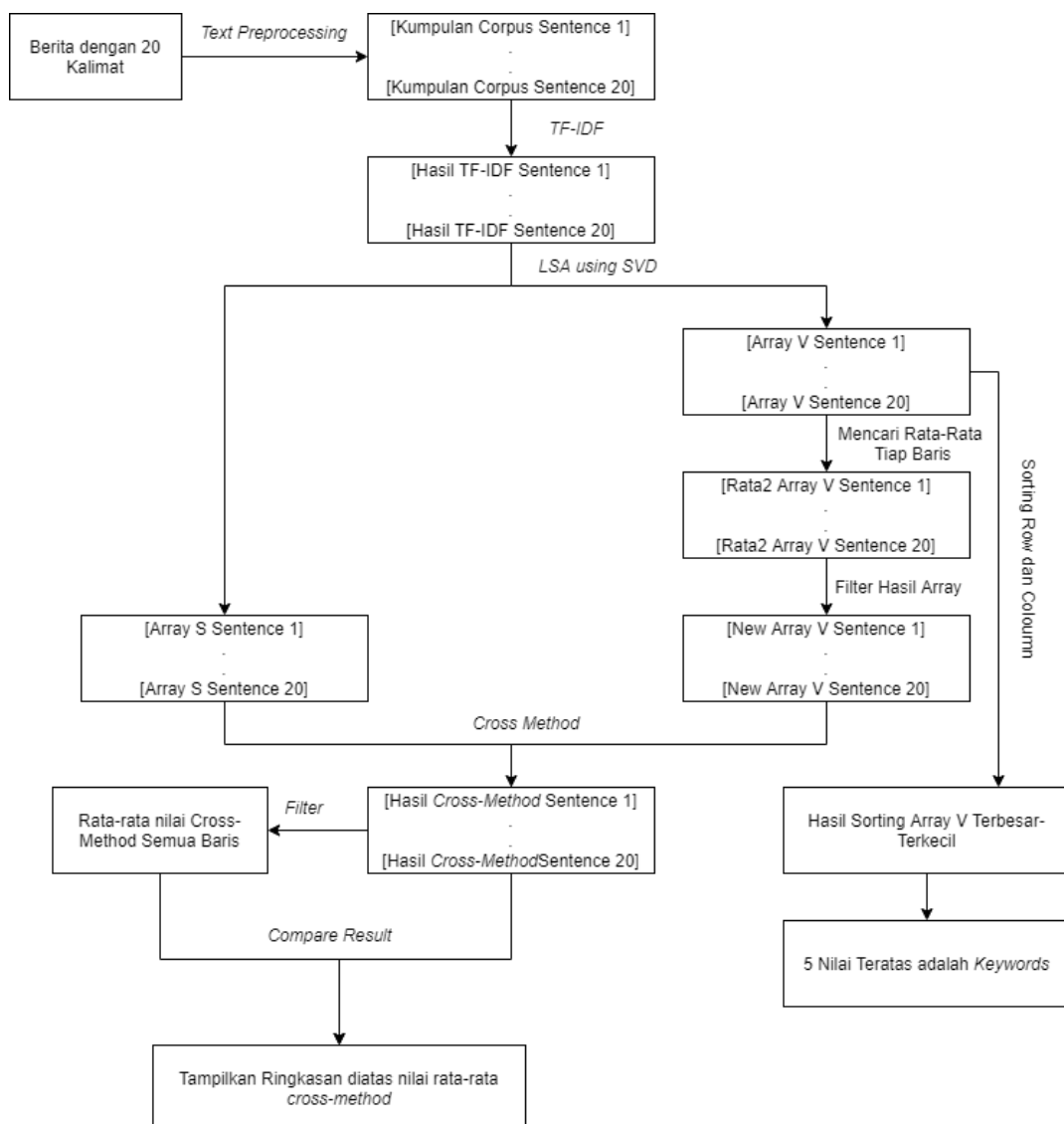
Langkah yang dilakukan adalah pertama mencari nilai rata-rata dari setiap baris matriks VT, setelah nilai rata-rata ditemukan akan dilakukan pencocokan dengan nilai setiap kata dari matriks VT, jika nilai kata lebih kecil dari nilai matriks maka nilai kata akan menjadi nol, sedangkan tidak nilai kata bernilai tetap. Jika kedua langkah tersebut telah dilakukan maka melakukan perhitungan dengan rumus:

$$length = \sqrt{\Sigma} = V^2 * S^2 \quad (4)$$

Keterangan:

- V : Matriks V Hasil SVD
- S : Matriks S Hasil SVD

Setelah perhitungan rumus, akan dilakukan perhitungan nilai rata-rata seluruh *length*, kalimat yang akan ditampilkan adalah kalimat yang mempunyai nilai lebih besar dari nilai rata-rata, juga mengambil kata-kata yang mempunyai bobot terbesar, lima bobot teratas akan dijadikan *keywords*, dimana *keywords* yang diambil mencakup topik yang dibahas dalam artikel yang telah diringkas, sehingga nantinya *keywords* diharapkan membantu perluasan pencarian dengan tidak terbatas dengan kata kunci masukan dan judul berita.



Gambar 5. Proses Pembentukan Ringkasan dan Keywords

3. HASIL PEMBAHASAN

Pengujian sistem dilakukan setelah tahapan mengkode selesai dilakukan, pengujian dalam penelitian ini dilakukan untuk menguji kemampuan fungsional sistem dari perangkat lunak.

Tabel 2 Tabel Pengujian *Compression Rate*

No	Dokumen	Compression Rate					
		RS(Jumlah Kalimat Ringkasan Sistem)	RM(Jumlah Kalimat Ringkasan Manual)	Kalimat Relevan	<i>F-measure</i>	<i>Recall</i>	<i>Precision</i>
1	1	13	13	12	92.3%	92.3%	92.3%
2	2	5	6	4	72.72%	66.7%	80%
3	3	6	7	4	61.54%	57.1%	66.7%
4	4	8	9	5	58.82%	55.56%	62.5%
5	5	8	8	6	75%	75%	75%
6	6	10	9	5	52.63%	55.56%	50%
7	7	7	9	7	87.5%	77.78	100%
8	8	3	4	3	85.71%	75%	100%
9	9	3	4	2	57.14%	50%	66.67%
10	10	4	5	4	88.88%	80%	100%
11	11	4	5	4	88.88%	80%	100%
12	12	7	7	6	85.71%	85.71%	85.71%

Pengujian ringkasan otomatis dilakukan dua kali dengan menggunakan *compression rate* sebesar 30% dari panjang artikel, dokumen yang diuji diambil *sample* dari setiap situs, dimana dokumen keseluruhan berjumlah seratus dua puluh, dan dokumen yang diuji diambil 10% dari dokumen keseluruhan, sehingga terdapat dua belas buah dokumen uji, hasil pengujian dapat dilihat pada tabel 1.

Hasil pengujian menghasilkan nilai *f-measure* dan *recall* terbesar 92.3% pada dokumen pertama, sedangkan nilai *precision* terbesar yaitu 100%. Pengujian ringkasan biasanya disebut dengan pengujian intrinsik, dimana *precision* dan *recall* digunakan untuk mengukur kualitas ringkasan sistem dengan cara membandingkan ringkasan sistem dengan ringkasan manual. *Precision* adalah tingkat ketepatan hasil ringkasan, sedangkan *recall* adalah tingkat keberhasilan ringkasan yang dihasilkan. Nilai *f-measure* digunakan untuk menghitung nilai akurasi ringkasan.

Pengujian *query* dilakukan dengan membuat berbagai macam modifikasi *query* untuk melihat kesesuaian dokumen yang ditampilkan dengan *query* masukkan, pengujian ini dapat dilihat melalui data riwayat pencarian yang terdapat pada sistem.

KESIMPULAN

Sistem yang dibuat mampu melakukan optimasi pencarian artikel Covid19 dari situs berita Kompas.com, Liputan6.com, Klikdokter.com, detikhealth.com dengan peringkasan teks otomatis menggunakan metode *latent semantic analysis*. Hasil pengujian menghasilkan nilai *f-measure* dan *recall* rata-rata sebesar 90.68% dan 85% dengan presentase data latih: data uji adalah 90:10. Pengumpulan data dilakukan selama bulan Februari-Juni 2020 diambil 200 dokumen latih, dan 20 dokumen uji. Pengujian dilakukan dengan *compression rate* sebesar 30%.

DAFTAR PUSTAKA

- Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S. M. M., & Williams, H. E. (2007). Stemming Indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing*. <https://doi.org/10.1145/1316457.1316459>
- Chakraborty Amartya, Sunanda Bose, (2020), Around the world in 60 days: an exploratory study of impact of COVID-19 on online global news sentiment, *Journal of Computational Social Science*.
- Doko, A., Štula, M., & Stipaničev, D. (2013). A recursive TF-ISF Based Sentence Retrieval Method with Local Context. *International Journal of Machine Learning and Computing*, 3(2), 195–200. <https://doi.org/10.7763/ijmlc.2013.v3.301>
- H Sheikha (2020). Text mining Twitter social media for Covid-19: Comparing latent semantic analysis and latent Dirichlet allocation, University of Gävle, Faculty of Engineering and Sustainable Development, Department of Computer and Geospatial Sciences. diva-portal.org
- Irmawati, I. (2017). Sistem Temu Kembali Informasi Pada Dokumen Dengan Metode Vector Space Model. *Jurnal Ilmiah FIFO*, 9(1), 74. <https://doi.org/10.22441/fifo.v9i1.1444>
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*. <https://doi.org/10.1080/01638539809545028>
- Larocca Neto, J., Santos, A. D., Kaestner, C. A. A., & Freitas, A. A. (2000). Generating text summaries through the relative importance of topics. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/3-540-44399-1_31
- L Singh, S Bansal, 2020 A first look at COVID-19 information and misinformation sharing on Twitter (2020), Cornell University, <https://arxiv.org/abs/2003.13907>, arxiv.org
- M Cinelli, W Quattrocioni, A Galeazzi (2020)- The covid-19 social media infodemic, Cornell University, arXiv preprint arXiv - arxiv.org
- Martin, C. D., & Porter, M. A. (2012). The extraordinary SVD. *American Mathematical Monthly*. <https://doi.org/10.4169/amer.math.monthly.119.10.838>
- Mustaqfiri, M., Abidin, Z., & Kusumawati, R. (2012). Peringkasan Teks Otomatis Berita Berbahasa Indonesia Menggunakan Metode Maximum Marginal Relevance. *Matics*. <https://doi.org/10.18860/mat.v0i0.1578>
- Shkapenyuk, V., & Suel, T. (2002). Design and implementation of a high-performance distributed web crawler. *Proceedings - International Conference on Data Engineering*. <https://doi.org/10.1109/ICDE.2002.994750>
- Steinberger, J., & Ježek, K. (2009). Evaluation measures for text summarization. *Computing and Informatics*.
- Vargiu, E., & Urru, M. (2012). Exploiting web scraping in a collaborative filtering- based approach to web advertising. *Artificial Intelligence Research*. <https://doi.org/10.5430/air.v2n1p44>