OTOMATISASI PERINGKASAN TEKS PADA DOKUMEN HUKUM MENGGUNAKAN METODE *LATENT SEMANTIC ANALYSIS*

TUGAS METODOLOGI PENELITIAN

Oleh:

MILLENIA RUSBANDI NIM. 1641720029

PROGRAM STUDI TEKNIK INFORMATIKA

JURUSAN TEKNOLOGI INFORMASI

POLITEKNIK NEGERI MALANG

2020

1. Tujuan

Berdasarkan latar belakang dan rumusan masalah di atas, maka tujuan dari skripsi ini yaitu:

1 Membantu pihak aparat hukum dalam melakukan peringkasan dokumen.
2 Menerapkan metode *latent semantic analysis* pada ringkasan dokumen hukum.
3 Membaca dokumen hukum dalam format pdf dan mengubahnya ke dalam teks yang dapat diolah oleh sistem peringkasan otomatis.
4 Menganalisis performasi hasil ringkasan dari metode tersebut berupa akurasi berdasarkan *precision*, *recall* dan *f-measure*.

2. Landasan Teori

Tinjauan pustaka merupakan bagian yang akan membahas tentang penyelesaian masalah yang akan memberikan jalan keluarnya. Dalam hal ini akan dikemukakan beberapa teori-teori yang berkaitan dengan masalah yang diangkat.

2.1 *State-of-the-Art* Penelitian Terdahulu

Berdasarkan penelitian pada tahun 2014, Agustinus Widiantoro telah membangun sebuah aplikasi peringkasan dokumen berbahasa jawa secara otomatis menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*). Dari penelitian tersebut, hasil yang didapatkan menunjukan bahwa tingkat keakurasian sistem mencapai 64%. Tetapi, kelemahan dari sistem ini yaitu kurangnya daftar kata umum (*stopwords*) yang digunakan sehingga penulis menyarankan untuk melakukan penambahan daftar kata umum untuk meningkatkan akurasi (Widiantoro, 2014).

Pada tahun 2015, penelitian dilakukan oleh Geetha J.K. dan Deepamala N. dengan judul "*Kannada text summarization using Latent Semantic Analysis*". Penelitian tersebut menggunakan Bahasa Kanada. Tingkat akurasi yang didapatkan yaitu 94% dan *precision* sebesar 80% (N & K, 2015).

Pada tahun 2017, penelitian dilakukan oleh Sohini Roy Chowdhury , Kamal Sarkar dan Santanu Dam dengan judul "*An Approach to Generic Bengali Text Summarization Using Latent Semantic Analysis*". Penelitian tersebut menggunakan

aksara Bengali. Teks yang diringkas diambil 10% dari teks asli. Nilai F-Score yang didapatkan yaitu 0.324347.

Pada tahun 2017, Huihong Lan dan Jinde Huang melakukan penelitian yang berjudul "*Chinese-English Cross-Lingual Text Clustering Algorithm based on Latent Semantic Analysis*". Klasterisasi menggunakan CLTC-LSA yang dilakukan, meningkat sebanyak 13.96% dibandingkan hanya menggunakan CLTC.

Pada tahun 2018, Mozibur Raheman Khan dan Rajkumar Kannan melakukan sebuah penelitian dengan judul "*Summarizing Health Review using Latent Semantic Analysis*". Penelitian ini menggunakan metode *Latent Semantic Analysis* untuk *feature identification*. Selanjutnya akan diklasifikasi menjadi *positive review* atau *negative review*. Tingkat akurasinya sebesar 82.20 %.

Penelitian berikutnya dilakukan oleh Tinaliah dan Triana Elizabeth pada tahun 2018. Peneliti menggunakan metode *Latent Semantic Analysis* dan *Jaro-Winkler Distance* dalam mendeteksi plagiarisme dokumen. Tingkat akurasi pada penelitian tersebut menghasilkan nilai plagiat mencapai 97,14% (Tinaliah & Elizabeth, 2018).

Dari penelitian yang telah dipaparkan diatas, Metode *Latent Semantic Analysis* dan Metode TF-IDF dapat diterapkan pada proses peringkasan dokumen dengan tingkat akurasi yang cukup baik. Maka dari itu, pada penelitian ini penulis menggunakan dua metode tersebut dalam melakukan peringkasan teks. Untuk objek yang akan digunakan yaitu berupa dokumen hukum.

Tabel 6.1 Tabel *State-of-the-Art* Penelitian Terdahulu

No	Judul	Penulis/Jurnal	Univ / Tahun	Permasalahan	Metode Pengolahan Data	Kesimpulan
1	Peringkasan Dokumen Berbahasa Jawa Secara Otomatis Menggunakan Metode TF-IDF	Agustinus Widiantoro	Universitas Sanata Dharma Yogyakarta / 2014	Pembuatan ringkasan masih manual membutuhkan waktu lama.	TF-IDF	Hasil yang didapatkan menunjukan bahwa tingkat keakurasian sistem mencapai 64%. Tetapi, kelemahan dari sistem ini yaitu kurangnya daftar kata umum (*stopwords*) yang digunakan.
2	*Kannada Text Summarization Using Latent Semantic Analysis*	*Geetha J.K. dan Deepamala N /* *International Conference on Advances in Computing,*	*RV College of Engineering* / 2015	Sulitnya melakukan peringkasan manual dengan adanya dokumen yang banyak.	*Latent Semantic Analysis*	Tingkat akurasi yang didapatkan yaitu 94% dan *precision* sebesar 80%

		Communications and Informatics (ICACCI)				
3	*An Approach to Generic Bengali Text Summarization Using Latent Semantic Analysis*	Sohini Roy Chowdhury , Kamal Sarkar dan Santanu Dam / *International Conference on Information Technology*	*Department of Computer Science and Engineering, Jadavpur University /* 2017	Sulitnya melakukan peringkasan manual dengan adanya dokumen yang banyak.	*Latent Semantic Analysis*	Penelitian tersebut menggunakan aksara Bengali. Teks yang diringkas diambil 10% dari teks asli. Nilai F-Score yang didapatkan yaitu 0.324347.
4	*Chinese - English Cross -Lingual Text Clustering Algorithm based on Latent Semantic Analysis*	Huihong Lan dan Jinde Huang / *Proceedings of Science ISSC*	*Guangxi College of Education Nanning, China* / 2017	Tingkat Akurasi dengan Menggunakan metode CLTC rendah.	*CLTC-LSA*	LSA mampu mengurangi *noise*. Klasterisasi meningkat sebanyak 13.96% dibandingkan hanya menggunakan CLTC.

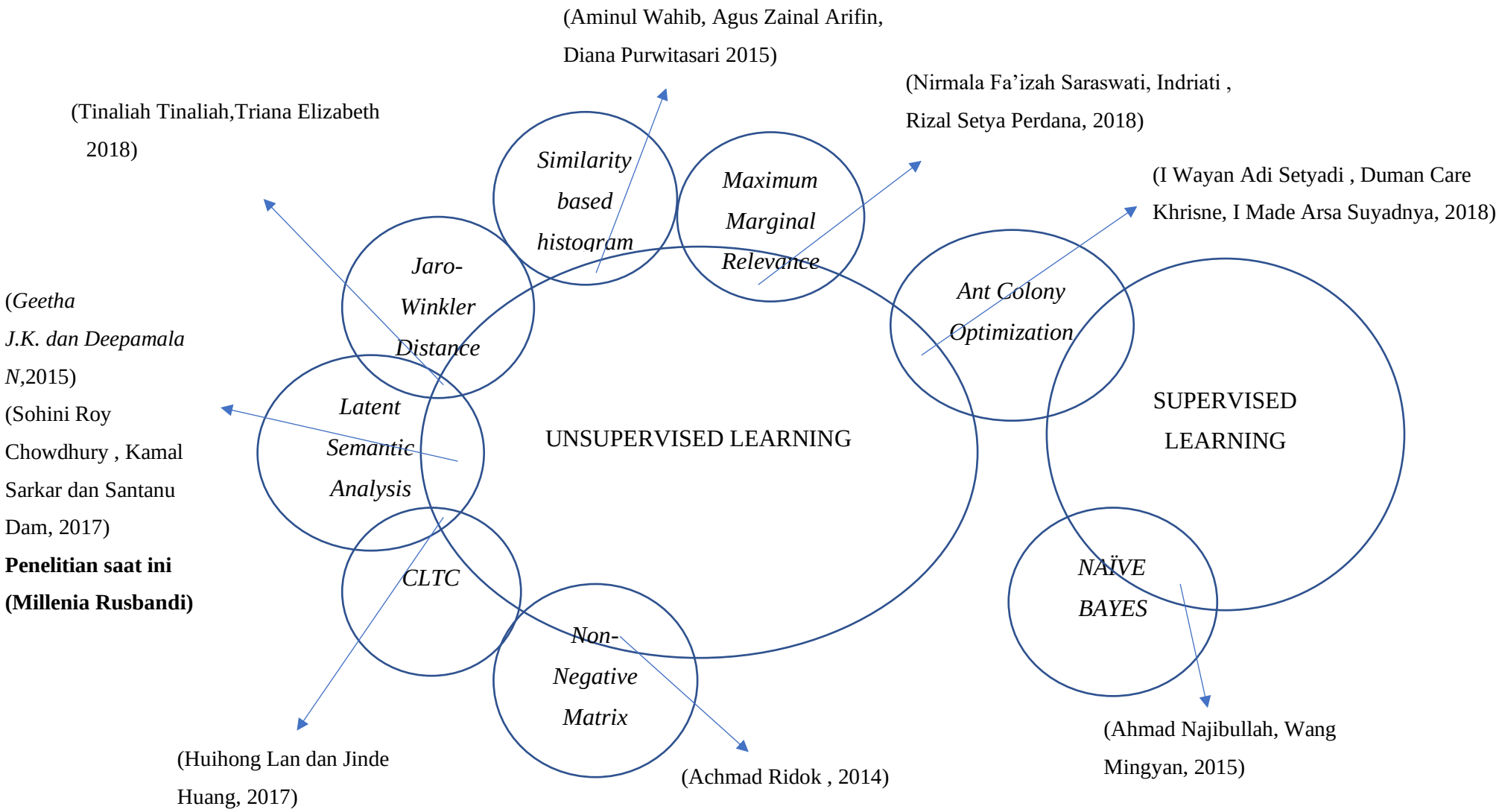
5	Peringkasan Dokumen Berbahasa Inggris Menggunakan Sebaran Local Sentence	Aminul Wahib, Agus Zainal Arifin, Diana Purwitasari / Jurnal Teknik Informatika ITS	Institut Teknologi Sepuluh November / 2015	Kurang efektif dalam mencari dan membaca informasi karena masih manual.	*Similarity based histogram clustering* (SHC), *Local Sentence*	metode sebaran *local sentence* lebih baik atau meningkat sebesar 13% dibandingkan dengan metode *SIDeKiCK*.
6	Perbandingan Hasil Deteksi Plagiarisme Dokumen dengan Metode *Jaro-Winkler* Distance dan Metode *Latent Semantic Analysis*	Tinaliah Tinaliah, Triana Elizabeth / Jurnal Teknologi dan Sistem Komputer; Volume 6, Issue 1, Year 2018	Akademi Manajemen dan Informatika MDP Palembang, STMIK Global Informatika MDP Palembang	Banyaknya plagiarisme.	*Latent Semantic Analysis* dan *Jaro-Winkler Distance*	Tingkat akurasi pada penelitian tersebut menghasilkan nilai plagiat mencapai 97,14%

			/ 2018			
7	Otomatisasi Peringkasan Dokumen Sebagai Pendukung Sistem Manajemen Surat	Ahmad Najibullah, Wang Mingyan / Register: Jurnal Ilmiah Teknologi Sistem Informasi, Januari 2015, Volume 1, Nomor 1	Universitas Nanchang, Nanchang, Republik Rakyat Tiongkok / *2015*	Pengelolaan surat tidak efektif dikarenakan banyaknya surat.	*Naïve Bayes*	Hasil uji coba menunjukkan bahwa tingkat kompresi adalah 53.67% dengan informasi penting yang tersedia dalam ringkasan mencapai 96.67% dari dokumen asli.
8.	*Automatic Text Summarization* Menggunakan Metode *Graph* dan Ant Colony Optimization	I Wayan Adi Setyadi , Duman Care Khrisne, I Made Arsa Suyadnya / Teknologi Elektro, Vol. 17, No.	*Universitas Udayana, Bali* / *2018*	Banyaknya dokumen tidak penting yang tersebar di internet, menyulitkan pencarian.	*Graph dan Ant Colony Optimization*	Pengujian hasil ringkasan dengan mencari kesamaan hasil ringkasan sistem dengan hasil ringkasan secara manual menggunakan cosine similarity memperoleh persentase kesamaan 76.3%. Pengujian

		1,Januari -April 2018				dengan autosummary tools pada Microsoft Word memperoleh hasil ringkasan ratarata memiliki kesamaan 68.15%. Hasil ringkasan sistem dengan hasil ringkasan ahli memiliki kesamaan 78.43%. Hal ini berarti lebih dari 75% informasi yang dianggap penting oleh manusia sudah dapat di temukan oleh sistem. 9

9.	Peringkasan Dokumen Bahasa Indonesia Berbasis *Non-Negative Matrix Factorization* (NMF)	Achmad Ridok / Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) Vol. 1, No. 1, April 2014, hlm. 39-44	Program Studi Ilmu Komputer, Universitas Brawijaya / 2014	Penggalian informasi dari dokumen berupa ringkasan secara otomatis masih tidak ada.	*Non-Negative Matrix Factorization (NMF)*	Hasil evaluasi menunjukkan ringkasan sistem mempunyai rata-rata *precision* dan *recall* masing-masing 0.19724 dan 0.34085. Sedangkan evaluasi ringkasan antar pakar mempunyai rata-rata presisi dan recall masing-masing 0.68667 dan 0.70642.
10.	Peringkasan Teks Otomatis Menggunakan Metode Maximum Marginal Relevance Pada Hasil Pencarian Sistem Temu Kembali Informasi Untuk Artikel Berbahasa Indonesia	Nirmala Fa'izah Saraswati, Indriati , Rizal Setya Perdana / Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer	Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya / 2018	Penggalian informasi dari dokumen berupa ringkasan secara otomatis masih tidak ada.	*Maximum Marginal Relevance* (MMR)	Hasil pengujian terbaik dari rata-rata precision, recall, f-measure dan akurasi masing-masing sebesar 0,70, 0,75, 0,70 dan 74,17. Metode yang

		e-ISSN: 2548-964X Vol. 2, No. 11, November 2018, hlm. 5494-5502				digunakan sudah cukup baik untuk mendapatkan dokumen yang relevan dengan query dan memperoleh ringkasan berdasarkan judul yang sesuai dengan isi dari dokumen.



Gambar 2.1 Posisi Penelitian

2.2 *Text Mining*

Text mining memiliki definisi menambang data yang berupa teks dimana sumber data di dapatkan dari dokumen. Tujuan *text mining* adalah mencari kata-kata yang dapat mewakili isi dari dokumen dan dilakukannya analisa keterhubungan antar dokumen.

Jenis masukan (*input*) dari *text mining* berupa data tak terstruktur yang merupakan pembeda utama dari *data mining* dimana menggunakan data terstruktur atau basis data sebagai masukan. *Text mining* dapat dianggap sebagai proses dua tahap yang diawali dengan penerapan struktur terhadap sumber data teks dan dilanjutkan dengan ekstraksi informasi dan pengetahuan yang relevan dari data teks terstruktur ini dengan menggunakan teknik dan alat yang sama dengan *data mining*. Proses yang umum dilakukan oleh text mining, yaitu perangkuman otomatis, kategorisasi dokumen, penggugusan teks, dll (Nindito, 2016).

2.3 Peringkasan Teks Otomatis

Peringkasan teks otomatis merupakan pembuatan rangkuman dari suatu teks secara otomatis dengan menggunakan serta memanfaatkan sistem peringkasan teks yang dijalankan pada komputer. Sebuah sistem peringkasan teks diberi *input* (masukan) berupa teks kemudian sistem akan memproses dengan melakukan peringkasan yang akan menghasilkan *output* (keluaran) berupa teks yang lebih singkat dari sumber teks aslinya (Hovy, 2005).Terdapat dua pendekatan peringkas teks yaitu:

a. Ekstraksi (*extractive summary*)

   Pada teknik ekstraksi, sistem menyalin unit-unit yang dianggap paling penting dari sebuah teks dan diubah menjadi ringkasan. Unit-unit teks yang disalin dapat berupa klausa utama, kalimat utama, atau paragraf utama tanpa ada penambahan kalimat-kalimat baru yang terdapat pada dokumen aslinya.

b. Abstraksi (*abstractive summary*)

   Teknik abstraksi menggunakan metode *linguistic* untuk memeriksa dan menafsirkan teks menjadi ringkasan. Ringkasan teks tersebut dihasilkan dengan cara menambahkan kalimat-kalimat baru yang merepresentasikan intisari teks

sumber ke dalam bentuk yang berbeda dengan kalimat-kalimat yang ada pada teks sumber (Gupta & Lehal, 2010).

Pada penelitian ini, metode yang digunakan untuk melakukan peringkasan teks otomatis adalah dengan menggunakan teknik ekstraksi. Hasil dari ringkasan merupakan kalimat asli yang terdapat pada dokumen dan tidak mengalami penambahan kalimat.

2.4 *Pre-Processing*

Pada tahapan ini, data tekstual akan diubah menjadi teks agar dapat diolah oleh sistem. Penelitian ini menggunakan dokumen sebagai inputan awal. Proses yang digunakan antara lain :

a. Pembentukan Kalimat

   Pembentukan kalimat yaitu pemecahan teks dokumen menjadi kumpulan kalimat berdasarkan delimiter.

b. *Case Folding*

   Case folding merupakan pengubahan huruf pada kalimat menjadi huruf kecil (*lowercase*) dan penghilangan karakter yang tidak valid seperti tanda baca.

c. *Tokenizing*

   Pada proses ini, kalimat tersebut dipecah kembali menjadi beberapa kata tunggal penyusunnya.

d. *Stopword Removal*

   Stopword removal adalah proses penghilangan kata-kata yang tidak merepresentasikan isi dokumen.

e. *Stemming*

   Stemming adalah proses pengembalian kata tunggal yang memiliki imbuhan menjadi kata dasar.

2.5 TF-IDF

Setelah dokumen diproses dengan cara *pre-processing, tokenizing, filtering* dan *stemming,* selanjutnya dilakukan proses pembobotan kata. Pada metode ini pembobotan kata dalam sebuah dokumen dilakukan dengan mengalikan nilai TF dan IDF.

Term frequency (TF) adalah pengukuran yang paling sederhana dalam metode pembobotan. Pada metode ini, masing-masing term diasumsikan mempunyai proporsi kepentingan sesuai jumlah kemunculan dalam teks dokumen. *Term frequency* dapat memperbaiki nilai *recall* pada *information retrieval*, tetapi tidak selalu memperbaiki nilai *precision* (Tokunaga & Iwayama, 1994). Hal ini disebabkan *term* yang frequent cenderung muncul di banyak teks, sehingga *term* tersebut memiliki kekuatan

Inverse document frequency (IDF) adalah metode pembobotan *term* yang lebih condong (fokus) untuk memperhatikan kemunculan *term* pada keseluruhan kumpulan teks. Pada IDF, term yang jarang muncul pada keseluruhan koleksi teks dinilai lebih berharga. Nilai kepentingan tiap *term* diasumsikan berbanding terbalik dengan jumlah teks yang mengandung *term* tersebut (Tokunaga & Iwayama, 1994).

Term frequncy inverse document frequency (TF•IDF) adalah metode pembobotan yang menggabungkan metode TF dan IDF. Metode ini diusulkan oleh Salton sebagai sebuah kombinasi metode yang dapat memberikan performansi yang lebih baik, khususnya dalam memperbaiki nilai *recall* dan *precision* (Tokunaga & Iwayama, 1994). Berikut ini merupakan perhitungannya :

$$TF.IDF = TF * \log(N/DF) \quad (7.1)$$

Keterangan :

TF : Jumlah *term* tersebut
N : Total dokumen
DF : Jumlah dokumen yang mengandung suatu *term*

2.6 Latent Semantic Analysis

Latent Semantic Analysis (LSA) menurut bahasa terbagi atas beberapa kata yang penting yaitu *latent* dan *semantic*, *latent* yang memiliki arti tersembunyi atau sesuatu yang masih belum terlihat, sedangkan semantic berasal dari bahasa yunani "*semanticos*" yang berarti memberi tanda, penting atau cabang linguistik yang mempelajari arti dan makna dari suatu bahasa, kode atau jenis representasi lainnya.

Dari pengertian dapat ditarik kesimpulan bahwa, LSA adalah menguraikan atau menganalisa makna yang masih tersembunyi dari suatu bahasa, kode atau jenis representasi lainya, guna memperoleh informasi yang penting. Kesamaan kata dan

kalimat diperoleh dengan cara menggunakan *Singular Value Decomposition* (SVD), di mana SVD mempunyai kapasitas untuk mereduksi *noise*, sehingga dapat meningkatkan hasil akurasi pada ringkasan (Peter & Kp, 2009).Peringkasan dokumen menggunakan metode LSA memiliki 3 tahapan yaitu : pembentukan matriks input dari dokumen untuk menampilkan kalkulasi, *Singular Value Decomposition*, dan penyeleksian kalimat.