PERINGKAS MULTI DOKUMEN MENGGUNAKAN METODE *K-MEANS* DAN *LATENT DIRICHLET ALLOCATION* (LDA) – *SIGNIFICANCE SENTENCES*

SKRIPSI

Disusun Sebagai Salah Satu Syarat
Untuk Memperoleh Gelar Sarjana Komputer
pada Departemen Ilmu Komputer/Informatika

Disusun Oleh:
SHIVA TWINANDILLA
24010314130078

DEPARTEMEN ILMU KOMPUTER/ INFORMATIKA
FAKULTAS SAINS DAN MATEMATIKA
UNIVERSITAS DIPONEGORO
2018

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Saya yang bertanda tangan di bawah ini :

Nama	:	Shiva Twinandilla
NIM	:	24010314130078
Judul	:	Peringkas Multi Dokumen Menggunakan Metode *K-Means* dan *Latent Dirichlet Allocation* (LDA) – *Significance Sentences*

Dengan ini saya menyatakan bahwa dalam skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan di dalam daftar pustaka.

Semarang, 18 Mei 2018

Shiva Twinandilla

NIM. 24010314130078

HALAMAN PENGESAHAN

Judul	:	Peringkas Multi Dokumen Menggunakan Metode *K-Means* dan *Latent Dirichlet Allocation* (LDA) – *Significance Sentences*
Nama	:	Shiva Twinandilla
NIM	:	24010314130078

Telah diujikan pada sidang skripsi pada tanggal 27 April 2018 dan dinyatakan lulus pada tanggal 27 April 2018.

Semarang, 18 Mei 2018

Mengetahui,
Ketua Departemen Ilmu Komputer/ Informatika

Dr. Retno Kusumaningrum, S.Si, M.Kom
NIP. 198104202005012001

Panitia Penguji Skripsi
Ketua,

Priyo Sidik Sasongko, S.Si, M.Kom
NIP. 197007051997021001

HALAMAN PENGESAHAN

Judul	:	Peringkas Multi Dokumen Menggunakan Metode *K-Means* dan *Latent Dirichlet Allocation* (LDA) – *Significance Sentences*
Nama	:	Shiva Twinandilla
NIM	:	24010314130078

Telah diujikan pada sidang skripsi pada tanggal 27 April 2018.

Semarang, 18 Mei 2018

Dosen Pembimbing

Dr. Retno Kusumaningrum, S.Si, M.Kom

NIP. 198104202005012001

ABSTRAK

Teknologi merupakan salah satu media yang digunakan untuk menyebarkan informasi ke khalayak umum. Di era globalisasi ini, ilmu pengetahuan dan teknologi terus berkembang pesat dari waktu ke waktu. Hal ini menyebabkan jumlah dokumen berita yang ada semakin banyak khususnya di internet. Dokumen berita *online* dapat membantu pembaca dalam memperoleh informasi terbaru secara cepat, dimanapun dan kapanpun. Namun, dokumen berita *online* mengesampingkan detail dan akurasi berita karena tujuannya untuk memberikan informasi terkini sebanyak-banyaknya. Banyak isi dokumen berita yang hampir sama sehingga menyebabkan redundansi dokumen berita atau disebut *yellow journalism. Yellow journalism* dapat menyebabkan pembaca sulit membedakan dokumen yang mengandung informasi fakta atau opini. Oleh sebab itu, diperlukan penelitian mengenai peringkas multi dokumen agar pembaca lebih mudah memahami maksud dari dokumen berita *online*. Peringkas multi dokumen menggunakan metode *K-Means* dan *Latent Dirichlet Allocation* (LDA) – *Significance Sentences* merupakan teknologi yang dapat diimplementasikan untuk mendapatkan hasil ringkasan dari beberapa dokumen berita yang secara umum memiliki topik yang sama. Tujuan dari penelitian ini yaitu untuk mengetahui kinerja metode peringkas multi dokumen menggunakan metode *K-Means* dan LDA – *Significance Sentences*. Pengujian sistem peringkas multi dokumen dilakukan dengan menggunakan metode ROUGE-1 dan terdapat 2 skenario pengujian. Pengujian pertama dilakukan untuk mengetahui nilai parameter terbaik pada metode LDA – *Significance Sentences.* Berdasarkan hasil pengujian pertama, penelitian ini memiliki nilai alfa terbaik sebesar 0.001 dengan nilai ROUGE-1 sebesar 0.5545 dan level peringkasan terbaik sebesar 30% dengan nilai ROUGE-1 sebesar 0.6118. Pengujian kedua dilakukan untuk mengetahui kinerja metode *K-Means* yang terdiri dari 2 proses dengan berita sebanyak 8 dokumen sehingga masing-masing proses menghasilkan 2 *cluster*. Proses pertama menghasilkan *cluster* 1 yang terdiri dari dokumen 1, 2, 3, 4, 6 dengan niai ROUGE-1 sebesar 0.6139 dan *cluster* 2 terdiri dari dokumen 5, 7, 8 dengan nilai ROUGE-1 sebesar 0.6199, sedangkan proses kedua menghasilkan *cluster* 1 yang terdiri dari dokumen 2 dengan nilai ROUGE-1 sebesar 0.5833 dan *cluster* 2 terdiri dari dokumen 1, 3, 4, 5, 6, 7, 8 dengan nilai ROUGE-1 sebesar 0.4542. Proses pertama memiliki hasil yang cukup baik karena nilai ROUGE-1 hampir mendekati nilai 1. Peringkas multi dokumen menggunakan metode K-Means dan LDA- *Significance Sentence* memiliki kinerja yang baik untuk metode LDA-*Significance Sentence* , sedangkan metode *K-Means* belum bisa membedakan dokumen berita berdasarkan topiknya secara khusus.

Kata kunci : Peringkas Multi Dokumen, berita *online, yellow journalism,* K-Means, *Latent Dirichlet Allocation, Significance Sentences,* ROUGE-1

ABSTRACT

Technology is one of the media used to disseminate information to the public. In this era of globalization, science and technology will continue to grow rapidly from time to time. This causes the number of existing news documents grew, especially on the internet. Online news documents can help readers to get the latest information quickly, wherever and whenever. However, online news documents override the details and accuracy of the news because of its purpose to provide up-to-date information as much as possible. Many of the contents news documents are almost the same that will led to redundancy of news documents or called yellow journalism. Yellow journalism can make it difficult for readers to distinguish documents containing fact or opinionated information. Therefore, it is necessary to extend more research about multi-document summarization so that readers can easily understand the intent of online news documents. Multi-document summarization using K-Means methods and Latent Dirichlet Allocation (LDA) - Significance Sentences is a technology that can be implemented to get a summary of some news documents that generally have the same topic. The purpose of this research is know the performance of multi-document summarization using K-Means method and LDA-Significance Sentences. Testing of multi-document summarization system is done using ROUGE-1 method and there are 2 test scenarios. The first test was conducted to find out the best parameter values in the LDA - Significance Sentences. Based on the first test result, this research has the best alpha value of 0.001 with ROUGE-1 value of 0.5545 and the best level of 30% with ROUGE-1 value of 0.6118. The second test is done to understand the performance of K-Means method consisting of 2 processes with news of 8 documents so that each process produce 2 cluster. The first process produces cluster 1 consisting of documents 1, 2, 3, 4, 6 with ROUGE-1 value of 0.6139 and cluster 2 consisting of 5, 7, 8 with ROUGE-1 value of 0.6199, while the second process produces cluster 1 which consists of document 2 with a ROUGE-1 value of 0.5833 and cluster 2 consists of documents 1, 3, 4, 5, 6, 7, 8 with a ROUGE-1 value of 0.4542. The first process has a pretty good result because the ROUGE-1 value is almost close to 1. Multi-document summarization using K-Means method and LDA-Significance Sentence has good performance for LDA-Significance Sentence, while K-Means method can not distinguish between news document by topic in particular.

Key Words : Multi-document summarization, online news, yellow journalism, K-Means, Latent Dirichlet Allocation, Significance Sentences, ROUGE-1

KATA PENGANTAR

Segala puji syukur bagi Allah SWT atas karunia-Nya yang diberikan kepada penulis sehingga penulis dapat menyelesaikan penulisan skripsi yang berjudul "Peringkas Multi Dokumen Menggunakan Metode *K-Means* dan *Latent Dirichlet Allocation* (LDA) – *Significance Sentences*". Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar sarjana strata satu pada Departemen Ilmu Komputer/Informatika Fakultas Sains dan Matematika Universitas Diponegoro Semarang.

Dalam penyusunan laporan ini banyak mendapat bimbingan dan bantuan dari berbagai pihak. Untuk itu, pada kesempatan ini penulis mengucapkan rasa hormat dan terima kasih kepada:

1. Dr. Retno Kusumaningrum, S.Si, M.Kom, selaku Ketua Departemen Ilmu Komputer/ Informatika FSM Universitas Diponegoro Semarang dan selaku dosen pembimbing yang telah membantu dalam proses bimbingan hingga selesainya skripsi ini.
2. Helmie Arif Wibawa, S.Si, M.Cs, selaku Koordinator Skripsi Departemen Ilmu Komputer/ Informatika FSM Universitas Diponegoro.
3. Mama Utiek, Papa Adjid, Kak Irla, Dek Dhiksa, Mas Taufan dan keluarga besar yang selalu memberikan motivasi dan doa dalam menyelesaikan skripsi ini.
4. Mbak Nisa, Dyosa, Nada, Putri, Enting, Bagas, Luthfi, Azys, Kharisma, Rayhan, Kak Dimas, Kak Wawan, Inna, Tiara, Rafa, Billa, Zaki, Widi, Jehan, Muti, Amrie, Nur Hakimah, Anjar, Annisa, Rona, Hanifah, Dino, Pandu, Johan dan teman-teman lainnya yang telah menghibur saat sedih, membantu, memberikan semangat dan doa kepada penulis dalam menyelesaikan skripsi ini.

Penulis menyadari bahwa laporan ini masih banyak kekurangan baik dari segi materi ataupun dalam penyajiannya karena keterbatasan kemampuan dan pengetahuan. Oleh karena itu, kritik dan saran sangan penulis harapkan. Semoga laporan ini bermanfaat bagi semuanya.

Semarang, 18 Mei 2018

Shiva Twinandilla

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR *SOURCE CODE*

DAFTAR LAMPIRAN

BAB I
PENDAHULUAN

Bab ini membahas mengenai latar belakang, rumusan masalah, tujuan dan manfaat, serta ruang lingkup pelaksanaan skripsi mengenai Peringkas Multi Dokumen Menggunakan Metode *K-Means* dan *Latent Dirichlet Allocation* (LDA) – *Significance Sentences.*

1.1. Latar Belakang

Di era globalisasi ini, ilmu pengetahuan dan teknologi terus berkembang pesat dari waktu ke waktu. Teknologi baru banyak bermunculan yang memiliki dampak positif atau negatif. Seiring berkembangnya jaman, banyak teknologi yang dapat digunakan untuk menyebarkan informasi. Hal itu menyebabkan jumlah dokumen berita yang ada semakin banyak khususnya di internet. Kini banyak masyarakat yang menggunakan internet untuk mempermudah segala urusannya. Hampir 60% masyarakat Indonesia menggunakan internet untuk mencari berita terbaru (Asosiasi Penyedia Jasa Internet Indonesia, 2015). Namun dokumen berita *online* memiliki kelebihan dan kekurangan. Kelebihannya yaitu *up to date*, mudah diakses dimanapun dan kapanpun. Sedangkan kekurangannya yaitu dokumen *online* mengesampingkan detail dan juga akurasi berita karena tujuannya untuk memberikan informasi terkini sebanyak-banyaknya kepada pembaca (Siregar, 2014). Kini banyak dokumen berita *online* yang memiliki isi dokumen hampir sama sehingga menyebabkan redundansi dokumen berita. Hal itu disebut *yellow journalism.* Efek negatif dari *yellow journalism* adalah pengguna sulit untuk membedakan dokumen yang ada merupakan informasi yang fakta atau tidak. Oleh karena itu, diperlukan penelitian mengenai peringkas multi dokumen berdasarkan kelompoknya agar lebih cepat memahami maksud dari beritanya.

Ringkasan merupakan teks yang dihasilkan dari satu atau lebih kalimat yang menyampaikan informasi penting dari dokumen (Verdianto, et al., 2016). Ringkasan dapat memudahkan pembaca dalam memahami tema dan konsep dalam dokumen, serta dapat mempersingkat waktu membaca. Terdapat dua metode meringkas dokumen yaitu metode

ekstraksi dan metode abstraksi (Chang & Chien, 2009). Metode ekstraksi merupakan metode meringkas dokumen dengan memilih bagian dari kata atau kalimat dalam dokumen asli. Sedangkan metode abstraksi merupakan metode meringkas dokumen dengan membuat kalimat baru yang memiliki informasi sama dengan dokumen asli.

Penelitian yang berkaitan dengan ringkasan multi dokumen Bahasa Inggris secara otomatis berbasis LDA sudah pernah dikerjakan. Arora dan Ravi (2008) menerapkan metode ekstraksi untuk melakukan peringkasan multi dokumen dengan menggunakan LDA dikombinasikan dengan model campuran untuk mengekstraksi topik dan membuat ringkasan dengan mengambil beberapa kalimat yang ada didalam dokumen tanpa memperhatikan detail tata bahasa serta struktur dokumen. Lukmana, dkk (2014) melakukan penelitian peringkasan multi dokumen dengan metode baru untuk merepresentasikan kalimat berdasarkan kata kunci dari topik teks menggunakan LDA. Penelitian tersebut mengelompokkan setiap kalimat ke dalam dokumen tertentu dengan menggunakan kesamaan histogram pengelompokkan (SHC) dan melakukan perangkingan *cluster* menggunakan *Sentence Information Density* (SID). Liu Na, dkk (2016) melakukan penelitian peringkasan multi dokumen dengan menggunakan pemilihan kalimat signifikan. Penelitian tersebut merupakan perbaikan dari penelitian sebelumnya yang menggunakan pendekatan topik signifikan untuk menghitung kesamaan antara topik degan kalimat dan dokumen (Na, et al., 2014). Penelitan terbarunya menggunakan LDA di awal lalu memperkenalkan 3 variabel yaitu distribusi kalimat (α), distribusi topik (β) dan kesamaan antara kalimat dengan judul dokumen (γ) untuk menentukan kalimat yang mempunyai bobot tinggi sebagai penyusun ringkasan (Na, et al., 2016). Kelebihan metode LDA yaitu cocok untuk data teks dalam jumlah besar (Liu, 2013) dan memiliki kinerja sangat baik dalam melakukan ekstraksi topik untuk dokumen teks berbahasa Indonesia (Prihatini, et al., 2017).

Akan tetapi untuk penelitian yang berkaitan dengan ringkasan multi dokumen Bahasa Indonesia secara otomatis masih sedikit yang mengerjakan. Hayatin, Fatichah, dan Purwitasari (2015) melakukan penelitian peringkasan multi dokumen dengan mempertimbangankan fitur penting berdasarkan *trending issue.* Fitur penting dalam berita yaitu *word frequency,* TF-IDF, posisi kalimat dan kemiripan kalimat terhadap judul (NeFTIS). Tahapan yang dilakukan yaitu ekstraksi *trending issue*, seleksi berita, ekstraksi

fitur berita, penghitungan total bobot kalimat dan penyusunan ringkasan (Hayatin, et al., 2015). Irawan, Hermawan, dan Samsuryadi (2016) melakukan penelitian peringkasan multi dokumen tanpa menghilangkan konten dan makna dari dokumen sehingga ringkasan tetap mengandung informasi yang dianggap penting. Penelitian tersebut menggabungkan metode *Latent semantic analysis* (LSA) dan metode *Maximum marginal relevance* (Irawan, et al., 2016). Verdianto, Arifin, dan Purwitasari (2016) melakukan penelitian peringkasan multi dokumen dengan pembobotan kalimat. Teknik pembobotan kalimat terbaik dengan menggunakan kombinasi keempat fitur yaitu *word frequency,* TF-IDF, posisi kalimat, dan kemiripan kalimat terhadap judul (Verdianto, et al., 2016).

Sementara peringkasan dokumen berbahasa Indonesia menggunakan LDA sudah dilakukan oleh Silvia, dkk pada tahun 2014. Silvia, dkk (2014) melakukan penelitian peringkasan untuk dokumen tunggal menggunakan *Latent Dirichlet Allocation* (LDA) dan algoritma genetika. Penelitian terdiri dari dua tahap yaitu tahap pelatihan dan pengujian. Pada tahap pelatihan digunakan untuk menghasilkan bobot fitur latih dari kalimat yang melibatkan proses membaca teks masukan, *presummarization, summarization* dan algoritma genetika. Sedangkan pada tahap pengujian digunakan untuk membuat ringkasan dari teks yang melibatkan proses membaca teks masukan, *presummarization, summarization* dan menyimpan ringkasan. *Presummarization* meliputi pemisahan konten dokumen teks kedalam paragraf, NLTK tokenizer untuk kalimat dan token kata, mengubah menjadi huruf kecil, *stopword removal*, dan *lemmatization* dengan kamus lookup ke dalam kamus Indonesia pada database MySQL. *Summarization* terdiri dari menghitung *Term Frequency- Inverse Sentence Frequency* (TS-ISF) bobot fitur, lokasi kalimat, dan panjang relatif dari kalimat, pemodelan topik LDA, kesamaan judul, kesamaan kata kunci, kalimat kohesi, dan data numerik. Untuk menghitung kesamaan judul, kesamaan kata kunci, dan kalimat kohesi menggunakan topik pemodelan dengan LDA dan *Jensen-Shannon Divergence*. Dari penelitian ini, LDA dan algoritma genetika sudah dapat menghasilkan ringkasan ekstraktif yang mencakup informasi penting dari dokumen teks tunggal lebih cepat (Silvia, 2014).

Akan tetapi, ringkasan dalam penelitian Silvia, dkk masih menghasilkan banyak dokumen ringkasan. Salah satu cara untuk mengurangi jumlah hasil ringkasan yang ada yaitu *clustering*. *Clustering* merupakan metode *unsupervised* yang mengkategorikan data

pada beberapa kelompok berdasarkan kesamaannya, maka tidak ada label untuk setiap *cluster* yang dihasilkan (Kusumaningrum & Farikhin, 2017). Salah satu metode yang dapat dipakai yaitu metode *K-Means*. Cukup banyak yang memakai metode *K-Means* karena mudah diimplementasikan. Penelitian yang berkaitan dengan metode *K-Means* sudah pernah dikerjakan, antara lain Clusterisasi Dokumen Web (Berita) Bahasa Indonesia Menggunakan Algoritma *K-Means* (Husni, et al., 2015), dan K-Means Algorithm Implementation For News Clustering (Larson, 2017). Dengan menggunakan algoritma K-Means, dokumen berita berhasil dikelompokkan secara otomatis sesuai dengan derajat kesamaan berita sehingga menjadi kelompok dokumen berita yang terstruktur (Husni, et al., 2015). Pada penelitian ini diberikan usulan *clustering* K-Means dan metode LDA-*Significance Sentences*. LDA-*Significance Sentences* memiliki performa yang bagus daripada algoritma *term frequency* (Na, et al., 2016).

1.2. Rumusan Masalah

Bagaimana membuat peringkas multi dokumen menggunakan metode *K-Means* dan *Latent Dirichlet Allocation* (LDA) – *Significance Sentences* untuk meringkas multi dokumen berbahasa Indonesia?

1.3. Tujuan dan Manfaat

Tujuan dari penelitian skripsi yaitu mengetahui kinerja metode peringkasan multi dokumen menggunakan metode *K-Means* dan LDA-*Significance Sentences.* Manfaat yang diharapkan dari penelitian skripsi ini yaitu sistem yang dibuat dan dikembangkan dapat memberikan kontribusi terhadap penelitian mengenai ringkasan multi dokumen agar dapat memudahkan manusia dalam memahami isi dari dokumen.

1.4. Ruang Lingkup

Ruang lingkup dalam menerapkan peringkas multi dokumen menggunakan metode *K-Means* dan LDA - *Significance Sentences* :

1. *Input* dari sistem ini berupa dokumen berita yang mencakup judul serta isi berita dalam bahasa Indonesia.

2. *Output* dari sistem ini berupa kalimat-kalimat yang ada di dalam isi berita berdasarkan bobot kalimat tertinggi.

1.5. Sistematika Penulisan

Sistematika penulisan yang digunakan dalam skripsi ini terbagi dalam beberapa pokok bahasan, yaitu :

BAB I PENDAHULUAN

Bab ini memberikan gambaran mengenai latar belakang masalah, rumusan masalah, tujuan dan manfaat, ruang lingkup serta sistematika penulisan skripsi mengenai peringkas multi dokumen menggunakan metode *K-Means* dan *Latent Dirichlet Allocation* (LDA) – *Significance Sentences.*

BAB II TINJAUAN PUSTAKA

Bab ini memberikan kajian pustakan yang berhubungan dengan tema skripsi sebagai landasan untuk perumusan dan analisis permasalahan pada skripsi. Kajian pustaka yang digunakan meliputi metode *K-Means, preprocessing,* LDA, *Sentence* LDA, *Significance Sentences,* ROUGE, dan Pengembangan Perangkat Lunak.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan mengenai tahapan dalam penyelesaian masalah skripsi. Tahapan tersebut meliputi *input* data, *preprocessing*, metode *K-Means*, LDA - *Significance Sentences*, ROUGE-1, analisa dan perancangan sistem.

BAB IV HASIL DAN ANALISA

Bab ini menguraikan hasil skenario eksperimen dan analisa pada penelitian yang dimulai dari teknis *input* data, penjelasan mengenai pengembangan sistem, semua skenario eksperimen dan analisa dari setiap hasil eksperimen yang telah dilakukan.

BAB V PENUTUP

Bab ini menjabarkan kesimpulan dari uraian yang telah diulas pada bab-bab sebelumnya dan saran untuk pengembangan penelitian lebih lanjut.