BIMASATI (Bulletin of Multi-Disciplinary Science and Applied Technology)
ISSN : 2809-6096 (Media Online)
Vol 1, No 3, April 2022
Hal : 100-104

Penerapan Algoritma Text Mining, Steaming Dan Texrank Dalam Peringkasan Bahasa Inggris

Leni Pertiwi

Fakultas Ilmu Komputer dan Teknologi Informasi, Program Studi Teknik Informatika, Universitas Budi Darma, Medan, Indonesia
Email: [1]lenipertiwi2003@gmail.com

Abstrak−Peringkasan teks dalam bahasa inggris gunakan untuk meringkas sebuah teks menggunakan komputer untuk mendapatkan ringkasan dari teks tersebut. Metode peringkasan teks dalam menggunakan ekstraktif dikarenakan metode ini mengambil informasi penting dari sebuah teks tanpa mengubah atau memodifikasi informasi tersebut. Salah satu algoritma yang dapat digunakan untuk meringkas teks dalam bahasa inggris adalah dengan menggunakan algoritma TextRank. Keunggulan dari algoritma TextRank yaitu tidak memerlukan pengetahuan mendalam tentang sebuah bahasa dan tidak memerlukan data latih untuk dapat meringkas teks. Cara kerja algoritma ini adalah dengan merepresentasikan kalimat pada teks ke dalam graf, lalu menghitung nilai tiap kalimat dengan menggunakan kesamaan (similarity) antar kalimat untuk menentukan hasil ringkasan. Selain menggunakan similarity untuk menentukan kalimat penting, pada penelitian ini juga menggunakan TextRank modifikasi yaitu dengan menggunakan levenshtein distance untuk menghitung ringkasan dengan membandingkan kemiripan antar string dengan cara memasukkan, menghapus, atau mengganti karakter string. Peringkasan teks dalam bahasa inggris menggunakan TextRank dilakukan dengan meringkas 100 teks berbahasa inggris yang selanjutnya akan dievaluasi menggunakan ROUGE. Evaluasi ROUGE bekerja dengan membandingkan hasil ringkasan dari TextRank dengan ringkasan manual oleh ahli pada bidang bahas inggris Untuk memudahkanperingakasan membutuhkan algoritma *text mining*, dengan menggunakan algoritma *text mining* dapat berguna untuk mendapatkan hasil sebenarnya.

Kata Kunci: Peringakasan; Bahasa Inggris; Text Mining; Textrank; Text

Abstract−Text summarization in English is used to summarize a text using a computer to get a summary of the text. The text summarization method uses extractives because this method takes important information from a text without changing it or the information. One of the algorithms that can be used to summarize text in English is by using the TextRank algorithm. The advantage of the TextRank algorithm is that it does not require in-depth knowledge of a language and does not require training data to be able to summarize text. The way this algorithm works is to represent sentences in the text into a graph, calculating the value of each sentence using questions (similarities) between sentences to determine the summary results. In addition to using similarity to determine important sentences, this study also uses a modified TextRank, namely by using levenshtein distance to calculate summaries by comparing the similarities between strings by entering, entering, or replacing character strings. Summarization of text in English using TextRank is done by summarizing 100 English texts which will then be evaluated using ROUGE. ROUGE evaluation works by comparing the summary results from TextRank with manual summaries by experts in the field of English. To facilitate the ranking requires a text mining algorithm, using text mining algorithms can be used to get actual results.

Keywords: Text Summarization; English; Text Mining; Textrank; Text

1. PENDAHULUAN

Jumlah dokumen digital mengalami peningkatan yang sangat pesat sehingga banyak memunculkan permasalahan baru dalam menggali dan memperoleh informasi secara cepat dan akurat. Jumlah dokumen yang terus berkembang menyebabkan para penggali informasi harus meluangkan waktu ekstra dalam mencari dan membaca informasi. Permasalahan lain yang muncul adalah besarnya potensi kehilangan informasi penting yang ada pada dokumen tersebut. Para peneliti mencoba menyelesaikan permasalahan ini dengan melakukan pengembangan metode dalam peringkasan dokumen dalam bahasa inggris.

Peringkasan dokumen adalah proses mengambil teks dari sebuah dokumen, menggali dan menyajikan informasi penting bagi user atau aplikasi dalam bentuk rangkuman yang singkat dan padat. Peringkasan dokumen dapat menjadi solusi bagi setiap orang yang tidak memiliki banyak waktu dan sedang membutuhkan informasi penting dalam tumpukan dokumen yang terus berkembang.

Orang yang tidak memiliki banyak waktu dan sedang membutuhkan informasi penting dalam tumpukan dokumen yang terus berkembang. Ringkasan dokumen yang baik adalah ringkasan yang mampu mencakup (coverage) sebanyak mungkin konsep-konsep penting (salient) yang ada pada dokumen sumber (Ouyang, dkk., 2013). Coverage dan saliency adalah masala utama dalam metode peringkasan dimana strategi pemilihan kalimat menjadi sangat penting karena harus mampu memilih kalimat-kalimat utama dan terhindar dari redundansi sehingga mampu mencakup banyak konsep (Suputra, dkk., 2013).

Penggunaan peringkasan teks dapat membantu memecahkan masalah ini. Dengan adanya peringkasan teks, diharapkan pembaca dapat dengan cepat dan mudah memahami makna sebuah teks tanpa harus membaca keseluruhan teks. Oleh karena itu, Automated Text Summarization (ATS) diperlukan dalam mengakomodasikan kebutuhan pengguna untuk mendapatkan hasil ringkasan teks dengan versi dokumen yang lebih kecil 50% atau kurang tetapi tetap berguna bagi pengguna Teknik yang umum digunakan dalam peringkasan teks adalah mengambil kalimat yang penting dari sebuah artikel tekstual tersebut. Hal ini dapat menghemat waktu pembaca karena dapat menghindari pembacaan teks yang tidak relevan dengan informasi yang diharapkan oleh pembaca, terutama ketika sangat banyak informasi tersedia di internet.

Text summarization terkait dengan natural language processing dan text mining. Beberapa proses yang terdapat dalam natural language processing dan text mining, juga digunakan dalam text summarization, tergantung metodenya. Terdapat dua buah pendekatan dilihat dari teknik pengambilan ringkasan yaitu ekstraksi (shallower approaches) dan abstraksi (deeper approaches). Pada teknik ekstraksi (shallower approaches), sistem menyalin informasi (kata atau kalimat) yang dianggap paling penting dari teks asli menjadi ringkasan (sebagai contoh, ku kalimat utama, atau paragraf utama). Sedangkan teknik abstraksi (deeper approaches), ringkasan yang menambahkan kata baru dan dapat merubah sususan kalimat. Pada umumnya, abstraksi dapat meringkas teks lebih kuat daripada ekstraksi, tetapi sistemnya lebih sulit dikembangkan karena mengaplikasikan teknologi natural language generation yang merupakan bahasan yang dikembangkan tersendiri.

Untuk itulah, peringkasan yang sering dikembangkan adalah peringkasan ekstraksi. Peringkasan berbasis graf merupakan suatu metode peringkasan teks yang dapat menghasilkan ringkasan ekstraktif. Metode berbasis graf termasuk sebuah pendekatan baru pada peringkasan teks. Walaupun pendekatan non-graf cukup berhasil menemukan unit teks yang paling penting dalam dokumen, teori graf dapat membantu pemahaman lebih baik terhadap keterhubungan antar unit teks. Teks sumber direpresentasikan menjadi sebuah graf sehingga disebut graf tekstual.

Text mining adalah salah satu teknik yang dapat digunakan untuk melakukan klasifikasi dimana, *text mining* merupakan variasi dari data mining yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar. Selain klasifikasi, *text mining* juga digunakan untuk menangani masalah clustering, information extraction, dan information retrival. *Text mining* telah menjadi lebih praktis bagi para ilmuwan data dan pengguna lain karena pengembangan platform data besar dan algoritma pembelajaran mendalam yang dapat menganalisi kumpulan data yang tidak terstruktur secara sebesar-besaran. Menganalisa teks membantu organisasi menemukan potensi wawasan bisnis yang berharga dalam dokumen perusahaan, email pelanggan, log call center, komentar survei verbatim, posting jaringan social, catatan medis dan sumber data berbasis teks lainnya. Semakin banyak, kemampuan penambangan teks juga dimasukkan ke dalam Al chatbots dan agen virtual yang digunakan perusahaan untuk memberikan tanggapan otomatis kepada pelanggan sebagai bagian dari pemasaran, penjualan, dan operasi layanan pelangganmereka. Penulis menerapkan algoritma *Text Mining* berdasarkan dari penelitian sebelumnya yang dilakukan oleh Setyo Budi. *Text Mining* Analisa Sentimen Review Flim Menggunakan Algoritma K-Means. Hasil pengujian menunjukkan bahwa accuracy algoritma K-Means dengan menggunakan dataset 300 dokumen review positif dan 300 dokumen review negative adalah 57.83%, dataset 700 dokumen positif dan 700 dokumen negative accuracy K-Means adalah 56.71% kemudian menggunakan dataset 1000 dokumen positif dan 1000 dokumennegative accuracy K-Means adalah 50.40%. Dan penelitian yang dilakukan oleh Budhi kurniawan wangsa, Darmawan Utomo, Saptadi Nugroho. Sistem Peringkas menggunakan *Generalized Vector Space Model:* Studi Kasus Berita diambil dari Media Massa Online. Dari hasil perancangan dan pengujian sistem peringkas berita ini dapat diambil sejumlah kesimpulan bahwa metode GVSM selain berfungsi untuk menilai keterkaitan antara topic per dokumen juga dapat di gunakan dalam menilai tingkat keterkaitan kalimat dengan topic suatu dokumen.

2. METODOLOGI PENELITIAN

2.1 Text Mining

Text Mining adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen. Tujuan dari text mining adalah mengekstrak informasi yang berguna dari sumber data.Jadi, sumber data yang digunakan pada text mining adalah sekumpulan dokumen yang memiliki format ,Tahapan text mining terdiri dariteks, Pengolahan teks (tokenisasi) adalah memecah kalimat menjadi kata per kata, perubahan huruf besar kehuruf kecil (kapitalisasi) dan menghilang kan tanda baca, Perubahan teks (stemming) adalah perubahan kata berimbuhan menjadi kata dasar, pemilahanteks (filtering) adalah melakukan perhitungan dan pengelompokkan kata per kata, Data Mining (Pattern Discovery) adalah proses pencarian pengetahuan atau pola yang menarik/bernilai, Evaluasia adalah penafsiran pola yangditemukan[2].

2.2 Algoritma Textrank

TextRank merupakan algoritma berbasis graf bekerja dengan cara memberikan peringkat pada teks dengan cara merepresentasikan objek dalam teks (Mihalcea & Tarau,2004). Dengan algoritma ini mengandalkan bentuk graf maka untuk dapat menentukan sebuahkalimat penting atautidak digunakan simpul (*node/verteks*). Dalam menerapkanalgoritma ini, langkahpertama yang dilakukan adalah membangun sebuah graf yangterdiri dari titik (verteks) mewakili tiap kalimat. Setiap kalimat akan dihubungkanberdasarkan kesamaan (*similarity)* antar kalimat biasa disebut dengan edge.

Algoritma TextRank merupakan algoritma yang bekerja dengan memberikan peringkat pada graf (Mihalcea &Tarau, 2004). Proses pada TextRank merepresentasikan isi dokumen ke dalam bentuk graf. Kalimat-kalimat tunggal dari proses sebelumnya dijadikan vertex. Nilai similarity setiap pasang kalimat dijadikan edge (Marsyah,2013). Menurut Mihalcea &Tarau (2004), Pada model TextRank untuk graf berbobot rumusnya didefinisikan sebagai berikut :

Dimana d sebagai damping factor yang bisa di beri nilai antara 0 dan 1. Untuk vertex Vi , In (Vi) di jadikan simpul yang mengarah ke simpul lainnya (predecessors) dan Out(Vi) di jadikan simpul yang mengarah ke vertex Vi (successors) sedangkan WS(Vi) merupakan jumlah skor simpul Vi. Menurut Mihalcea &Tarau (2004), implementasi algoritma yang berbasis graf pada pemrosesan text bahasa alami memiliki 4 tahapan yaitu:

Pada dasarnya proses kerja dari text mining banyak mengapdopsi dari penelitian data mining namun yang menjadi perbedaan adalah pola yang digunakan oleh text mining diambil dari sekumpulan bahasa alami yang tidak terstruktur sedangkan dalam data mining pola yang diambil dari database yang terstruktur (Han & Kamber, 2006).

1. Identifikasi unit teks yang paling sesuai untuk dijadikan simpul padagraf
2. Pemberian sisi antar simpul unit teks baik dengan bobot atau tidak berarahatau tidakberarah.
3. Proses menggunakan algoritma hingga objek graf bertemu satu dengan lainnya (convergence).
4. Urutkan simpul berdasarkan skornya. Nilai skor selanjutnya dimasukkan kedalam simpul pada saat prosesalgoritma.

Text Rank efektif diaplikasikan pada peringkasan dokumen dikarenakan tidak memerlukan data tambahan dan pengetahuan spesifik bahasa pada dokumen yang akan diringkas. Algoritma TextRank juga tidak membutuhkan data training dan bisa digunakan dalam semua bahasa pemrograman (Mihalcea dan Tarau 2004).

$$\mathrm{WS}(\mathrm{V}i) = (1 - \mathrm{d}) + \mathrm{d} \times \Sigma \mathrm{vj} \in \ln(vi) \frac{Wji}{\Sigma vk \epsilon\, out(vj) wjk} WS\,(vj) \qquad (1)$$

3. HASIL DAN PEMBAHASAN

3.1 Analisa Masalah

Analisa merupakan suatu kegiatan atau usaha dalam mengamati suatu hal atau benda dengan cara menguraikan suatu keseluruhan menjadi suatu komponen untuk memecahkan suatu masalah sehingga dapat mengenal tanda-tanda hubungan satu sama lain dan fungsi masing-masing dalam satu keseluruhan yang terpadu kemudian dicari kaitannya lalu ditafsirkan maknanya.

Ringkasan yang dibutuhkan untuk mendapatkan isi berita secara ringkas. Konsep sederahana dalam mengambil bagian penting dari keseluruhan isi berita kemudian menyajikannya kembali dalam bentuk yang lebih singkat. Peringkasan teks yang akan dibuat merupakan sistem yang secara ringkas dapat membaca teks single dokumen dan akan menghasilkan sebuah ringkasan. Metode yang digunakan dalam peringkasan menggunakan pendekatan ekstraksi yaitu algoritma Text Miing Steaming, Text Rank

Proses secara umum dalam pembuatan ringkasan otomatis pada skripsi ini, yaitu text preprosessing, meliputi pemecahan kalimat, case folding, filtering, tokenizing kata dan stemming. Ketika teks akan diringkas, proses yang dilakukan adalah:

1. User memasukkan teks dokumen yang akan diringkas dan memasukkan kalimat query
2. Kemudian sistem melakukan pemrosesan teks (text preprosessing), yaitu pemecahan kalimat, case folding, filtering, tokenizing kata dan stemming.

3.1.1 Penerapan Algoritma TextRank

Untuk meringakas maka memerlukan data dokumen, penulis mengambil data tentang berita yang ingin diringkas. Data di ambil secara *real- time* dari website liputan6.com. Pengguna (*reader*) dapat membaca berita kriminal secara ringkas dengan menyertakan poin dan unsur penting dalam sebuah berita. Pengguna juga dapat menentukan berapa jumlah baris ringkasan yang akan ditampilkan dalam aplikasi. Untuk memudahkan dalam melakukam analisa, maka penulis telah menentukan judul berita **"Nepal Rescuers Find 3 Bodies Near Crashed US Marine"** adapun isi berita dari website liputan6.com sebagai berikut

o **"Nepal Rescuers Find 3 Bodies Near Crashed US Marine"**

Nepalese rescuers on Friday found three bodies near the wreckage of a U.S. Marine helicopter that disappeared earlier this week while on a relief mission in the earthquake-hit Himalayan nation, fficials said. Nepal's Defense Secretary IsworiPoudyal gave no details about the nationalities of the victims. The helicopter was carrying six Marines and two Nepalese army soldiers. The wreckage was found near Gothali village in the district of Dolakha. The U.S. Embassy in Nepal had no immediate comment Friday.

The discovery of the wreckage, first spotted by a Nepalese army helicopter Friday, followed days of intense search involving U.S. and Nepalese aircraft and even U.S. satellites. The U.S. relief mission was deployed soon after a magnitude-7.8 quake hit April 25, killing more than 8,200 people. It was followed by another magnitude-7.3 quake on Tuesday that killed 117 people and injured 2,800. The second quake was centered between Kathmandu and Mount Everest, and hit hardest in deeply rural parts of the Himalayan foothills, hammering many villages reached only by hiking trails and causing road-blocking landslides. (Source: Nepal rescue team finds three bodies near crashed U.S. Marine helicopter, The Globe and Mail, May 15 2015)

Gambar 1. Berita Liputan6.com

Pada aktivitas yang dilakukan dalam proses ini adalah memecah string dokumen utuh menjadi kalimat-kalimat dengan menghilangkan delimiter atau tanda baca yang menyusunnya seperti titik ".", tanda tanya "?", dan tanda seru "!".

Tabel 1. Pemecahan Kalimat

NO	Kalimat
1	On Friday, Nepalese rescuers found three bodies near the wreckage of a United States Marine helicopter
2	who went missing earlier this week while on an aid mission in the earthquake-hit Himalayan country
3	according to the authorities. Nepal's Minister of Defense, IsworiPoudyal
4	did not provide details on the nationalities of the victims
5	The helicopter is carrying six Marines and two Nepalese soldiers. The wreckage of the plane

Dokumen yang telah dipotong menjadi beberapa kalimat kemudian di tahap *case folding* ini yang selanjutnya dilakukan yaitu mengubah teks menjadi huruf kecil, menghilangkan angka dan tanda baca maupun simbo-simbol karena

Tabel 2. Hasil case Folding

NO	Kalimat
1	The wreckage was found near the village of Gothali in the Dolakha district.
2	he great United States in Nepal did not immediately comment
3	these events on Friday.
4	The ruins were first discovered by an army helicopter

Pada tahapan ini akan dilakukan adalah mengambil kata-kata penting dari hasil case folding kalimat dan membuang kata-kata yang dianggap kurang penting. Algoritma yang dipakai adalah *stoplist* (membuang kata yang kurang penting) atau *wordlist* (menyimpan katapenting).

CStoplist (stopword) adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag- of-words*. Dalam tahap pembuangan kata-kata yang tidak penting adalah kata hasil parsing dicek dengan kamus (kumpulan kata) *stopword.* Jika kata parsing ada yang sama dengan stopword maka kata akan dibuang atau dihapus. *Stopword* yang digunakan dalam penelitian ini dapat dilihat pada halamanlampiran.

Tabel 3. Hasil *Filtering* kata

No	Kalimat
1	Nepal on Friday, followed by intense searches involving aircraft
2	United States and Nepal, even using United States satellites.
3	United States aid mission deployed after a magnitude earthquake
4	7. on the Richter scale struck on April 25 which killed more than 8,200 people
5	This was followed by an aftershock measuring 7.3 on the Richter scale on Tuesday which killed 117 people and injured 2,800.
6	e second earthquake was centered between Kathmandu and Mount Everest.

Tahap tokenizing adalah tahap pemotongan string menjadi potongan kata kemudian disusun menjadi baris. Pemotongan string kalimat-kalimat hasi *filtering* berdasarkan delimiter yang menyusunnya yaitu karakter spasi ("").

Tabel 4. Hasil *Tokenzing* kata

No	Kata	No	Kata	No	Kata	No	Kata
1	Numbers	5	Effirot	9	Corpase	13	To do
2	Spice	6	Fide	10	Contion	14	Help
3	Lost	7	Will	11	Search	15	Worride
4	Diffcult	8	Not	12	Rescue	16	Pleane

4. KESIMPULAN

Berdasarkan hasil rancangan sistem yang dilakukan oleh penulis tentang peringkasan teks dalam bahasa inggris dalam penyusunan ini maka dapat diambil kesimpulan yang merupakan hasil akhir Dengan diterapkan beberapa proses pada preprocessing seperti, Case folding,tokenizing,filtering, dan Stemming juga proses dalam pemberian bobot pada tiap kata dan teknik reduksi dimensi untuk mengurangi dimensi jumlah kata dapat membantu perhitungan probabilitas kata dan kecepatan pada proses peringkasan kata dalam bahasa inggris. Jumlah kata yang dihasilkan dari peringkasan sangat mempengaruhi hasil akurasi pada proses peringkasan dari jumlah awal 4500 kata mampu di kurangi hingga 327 kata.Adanya nilai bobot dari suatu term dapat memudahkan paragraf mana yang akan diawali untuk diringkas.

REFERENCES

[1] D. Nurnaningsih and A. A. Permana, "Rancangan Aplikasi Pengamanan Data Dengan Algoritma Advanced Encyption Standard (Aes)," *J. Tek. Inform.*, vol. 11, no. 2, pp. 177–186, 2018, doi: 10.15408/jti.v11i2.7811.

[2] P. Soepomo, "Penerapan Text Mining Pada Sistem Klasifikasi Email Spam Menggunakan Naive Bayes," *Penerapan Text Min. Pada Sist. Klasifikasi Email Spam Menggunakan Naive Bayes*, vol. 2, no. 3, pp. 73–83, 2014, doi: 10.12928/jstie.v2i3.2877.

[3] I. WARMAN and R. RAMDANIANSYAH, "ANALISIS PERBANDINGAN KINERJA QUERY DATABASE MANAGEMENT SYSTEM (DBMS) ANTARA MySQL 5.7.16 DAN MARIADB 10.1," *J. Teknoif*, vol. 6, no. 1, pp. 32–41, 2018, doi: 10.21063/jtif.2018.v6.1.32-41.

[4] Defta Afriani, "Perancangan Knowledge Management System dengan SECI Model Pada Layanan Perbaikan AC Mobil di Bengkel Agung Motor Cinere Menggunakan VB.NET," *Inform. SIMANTIK*, vol. 4, no. 1, pp. 29–35, 2019.

[5] R. Melita *et al.*, "(TF-IDF) DAN COSINE SIMILARITY PADA SISTEM TEMU KEMBALI INFORMASI UNTUK MENGETAHUI SYARAH HADITS BERBASIS WEB (STUDI KASUS : SYARAH UMDATIL AHKAM)," vol. 11, no. 2, 2018.

[6] D. Andriani and M. T. Furqon, "Peringkasan Teks Otomatis Pada Artikel Berita Hiburan Berbahasa Indonesia Menggunakan Metode BM25," vol. 3, no. 3, pp. 2603–2610, 2019.

[7] Suendri, "Implementasi Diagram UML (Unified Modelling Language) Pada Perancangan Sistem Informasi Remunerasi Dosen Dengan Database Oracle (Studi Kasus: UIN Sumatera Utara Medan)," *J. Ilmu Komput. dan Inform.*, vol. 3, no. 1, pp. 1–9, 2018, [Online]. Available: http://jurnal.uinsu.ac.id/index.php/algoritma/article/download/3148/1871.

[8] Y. Heriyanto, "Perancangan Sistem Informasi Rental Mobil Berbasis Web Pada PT.APM Rent Car," *J. Intra-Tech*, vol. 2, no. 2, pp. 64–77, 2018.

[9] E. Z. Henry Februariyanti, "Rancang Bangun Sistem Perpustakaan untuk Jurnal Elektronik," *J. Teknol. Inf. Din.*, vol. 17, no. 2, pp. 124–132, 2012.

[10] M. P. Simatupang and D. P. Utomo, "Analisa Testimonial Dengan Menggunakan Algoritma Text Mining Dan Term Frequency-Inverse Document Frequence (Tf-Idf) Pada Toko Allmeeart," KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer), vol. 3, no. 1, pp. 808–814, 2019.