Available *online*at : *http://bit.ly/InfoTekJar*

InfoTekJar :Jurnal Nasional InformatikadanTeknologiJaringan

ISSN (Print) 2540-7597|ISSN (Online) 2540-7600

Sistem cerdas

Peringkasan paper dengan metode Sparse Nonnegative Matrix Factorization untuk Pemeriksaan Kesesuaian dengan Abstrak Tugas Akhir

Irwan Darmawan[1], Reddy Alexandro H[2], Hendrawan Armato[2]

[1]*Universitas Madura, jl. Panglegur K. 35, Pamekasan, 325786, Indonesia*
[2]*Institut Sains dan Terapan Teknologi Surabaya, jl. Ngagel Jaya Tengah 73-77, Indonesia.*

KEYWORDS

Perangkuman Otomatis, SNMF, Bobot Graf, Recall.

CORRESPONDENCE

Phone: +6285231847739

E-mail: darmawan@unira.ac.id

A B S T R A C T

Salah satu jurusan yang terdapat di Universitas Madura adalah teknik informatika dimana ketika mahasiswa sebelum yudisium maka harus mengumpulkan paper terlebih dahulu sebagai salah satu syarat untuk yudisium. Isi dari paper yang tulis oleh mahasiswa terkadang tidak sesuai dengan abstrak paper tersebut. Sehingga diperlukan sebuah sistem yang dapat memeriksa kesesuaian isi dari paper dengan abstrak paper yang ditulis.

Metode yang digunakan dalam meringkas dokumen/paper adalah dengan menggunakan model graph. Model graph ini digunakan untuk menentukan bobot masing-masing kalimat supaya dapat di cluster. Cluster yang digunakan adalah dengan menggunakan metode SNMF (Sparse Non Negative Matrix Faktorization) kemudian dari hasil cluster diambil 2 kalimat tertinggi dari hasil masing-masing cluster untuk menentukan hasil ringkasan, dan sebagai pembandingnya adalah hasil ringkasan yang dibuat oleh pakar.

Untuk membuktikan seberapa efektif metode SNMF dalam menyelesaikan permasalahan, penulis melakukan beberapa uji coba. Setelah melewati beberapa uji coba, penulis menyimpulkan bahwa algoritma SNMF mampu menyelesaikan dengan baik untuk permasalahan pada peringkasan paper dan mampu memeriksa kesesuaian abstrak terhadap tugas akhir mahasiswa Universitas Madura dengan hasil akurasi recall 54.63.

PENDAHULUAN

Tugas akhir merupakan karya ilmiah yang dihasilkan mahasiswa pada sebuah perguruan tinggi dimana tugas akhir tersebut dibuat dalam bentuk laporan akhir. Di Universitas Madura jika seorang mahasiswa sudah memenuhi syarat sks (satuan kredit semester) untuk mencapai kelulusan maka mahasiswa tersebut harus menyusun tugas akhir atau skripsi terlebih dahulu kemudian membuat paper tugas akhir sebagai salah satu syarat untuk memperoleh gelar sarjana.. Hal inilah yang mendasari penulis untuk membuat sistem yang dapat mem-validasi kesesuain antara isi dan abstrak paper tugas akhir yang dibuat oleh mahasiswa Universitas Madura.

Pada dasarnya penggunaan peringkasan pada dokumen tunggal dimaksudkan untuk mempermudah para pembaca untuk mengambil kesimpulan dari isi sebuah dokumen. Pada perkembangan berikutnya peringkasan dokumen digunakan salah satunya untuk sistem checker kesesuaian abstrak yang ditulis dengan isi dokumen tugas akhir mahasiswa pada program studi Teknik Informatika Universitas madura sehingga mempermudah mahasiswa untuk menyesuaikan abstrak yang ditulis. Pendekatan metode yang digunakan dalam sistem ini adalah sparse non negative matrik faktorization untuk cluster sebuah dokumen tunggal. Hasil dari cluster tersebut digunakan sebagai data latih untuk diambil sebagai acuan batasan nilai ambang terbaik terbaik dari sebuah paper mahasiswa yang akan diterima atau ditolak oleh sistem sehingga dengan batasan nilai ambang tersebut mahasiswa yang memasukkan isi papernya dan abstrak tersebut secara otomatis dapat mengetahui hasilnya berapa prosentase kesesuaian abstraknya terhadap paper yang di inputkan kedalam sistem peringkasan sehingga mahasiswa dapat menulis ulang abstrak yang tidak sesuai tersebut.

Meringkas berarti mengambil inti sari isi dari sebuah dokumen baik dokumen tunggal maupun multi dokumen. Pada penelitian ini dikhususkan untuk meringkas pada dokumen tunggal dimana sistem pemodelan peringkasan dalam garis besarnya dibagi menjadi tiga pedekatan peringkasan dokumen yaitu abstraktif, ektraktif dan compressive summarization. Pada pemodelan sistem ini menggunakan pendekatan ektraktif yaitu ringkasan yang memotong kata atau kalimat dari sebuah dokumen aslinya kemudian menyusunnya kembali sehingga menghasilkan sebuah ringkasan dokumen. Dalam peringkasan dokumen yang akan dilkukan juga terdapat beberapa bagian penting yaitu perangkingan sebuah kalimat, kalimat yang di dapat dari hasil ektraksi sebuah dokumen dirangking menggunakan metode

https://doi.org/10.30743/infotekjar.v5i1.2473

iterasi sebuah rangking kalimat sehingga didapatkan kalimat mana yang paling banyak dirujuk oleh kalimat yang lain tujuannya adalah untuk megurutkan kalimat yang paling dianggap penting berdasarkan metode parangkingan kalimat.

Dalam meng-cluster isi dari sebuah dokumen ada banyak metode yang digunakan tetapi dalam paper ini yang digunakan adalah SNMF, pada perkembangan awal SNMF adalah metode yang berasal dari metode NMF (Non Negative matrix faktorization) dimana metode ini mengektraksi dari objek yang sangat besar menjadi beberapa bagian dengan kombinasi linier non negative. Cara kerja metode ini adalah memecah dokumen teks ke dalam kalimat-kalimat dan menghitung frekuensi masing-masing term dalam kalimat yang digambarkan dengan matrik nonnegative A berukuran mxn, m adalah jumlah kata dan n jumlah kalimat dalam dokumen. Matrik A didekomposisi ke dalam suatu perkalian matrik fitur semantik berukuran mxr W dan matrik variabel semantik tidak negativeberukuran rxn H. Nilai r dipilih lebih kecil dari m atau n sehingga total ukuran W dan H lebih kecil dari matrik A. Kemudian selanjutnya terbentuk cluster berdasarkan matrik W dan H tersebut dimana matrik W dan H tidak bernilai negative pada tahapan selanjutnya akan di iterasi sampai mencapai nilai tertentu berdasarkan banyaknya iterasi yang ditentukan terlebih dahulu. Dari hasil iterasi akan ketemu pengurutan bobot kalimat yang paling penting terhadap kalimat yang lain kemudian dipilih kalimat yang masuk ke cluster tertentu yang sudah ditentukan banyaknya cluster yang di inginkan.

Dalam membandingkan kalimat pada abstrak dan isi dokumen menggunakan sistem rouge SR-SNMF(sentence rangking – sparse non negative faktorization). ROUGE merupakan suatu teknik evaluasi yang digunakan untuk menentukan kualitas dari sebuah ringkasan secara automatis dengan cara membandingkan ringkasan yang dihasilkan oleh ATS (Automatic text summarization) dengan ringkasan (ideal) lain yang dibuat oleh manusia.

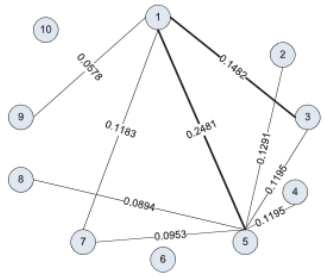
Penelitian ini bertujuan untuk memanfaatkan hasil dari peringkasan teks atau dokumen tunggal untuk dibuat perbandingan tingkat kemiripan abstrak terhadap isi dari dokumen tersebut pada laporan tugas akhir mahasiswa Universitas Madura. Jika tingkat kemiripannya memenuhi nilai batas yang ditentukan maka isi dari abstrak mahasiswa tersebut diterima oleh sistem sebaliknya jika tidak memenuhi nilai batas yang telah ditentukan maka mahasiswa tersebut direkomendasikan untuk menulis ulang abstrak tersebut.

METHOD

Dalam melakukan penelitian ini , terdapat beberapa tinjauan pustaka yang digunakan sebagai referensi rujukan sebagai dasar teori yang digunakan untuk membantu menyelesaikan permasalahan yang akan dibahas. Tinjauan pustaka tersebut dapat dikaji sebagai berikut.

A. Pembobotan Model Graf dan rangking Kalimat

Metode ini diperkenalkan oleh Shuzi Sam Ge, Zhengchen Zang, Hongsheng He Journal IEEE 20131. Dalam peringkasan sebuah dokumen, algoritma yang umum digunakan adalah algoritma yang merangkum sebuah dokumen tanpa menambahkan informasi apapun. Jenis algoritma yang lain adalah melakukan perangkuman dari dokumen berdasarkan topik atau query yang diberikan.



Gambar. 1. Ilustrasi Graph Model

Pada gambar graph diatas dapat dilihat contoh hasil pembobotan kalimat yang dimodelkan dengan pemodelan graph, pada nomor 10 tidak memiliki link pada kalimat manapun maka bobot kalimat tersebut adalah berbeda dengan yang lain yang memiliki bobot antar kalimat maka nilainya tidak sama dengan 0(nol). Bila terdapat kemiripan antar dokumen melebihi 0 (nol) maka dua bagian akan dihubungkan dengan sebuah bobot antar keduanya yaitu antar kalimat ***Si*** dan kalimat ***Sj*** Rumus persamaannya adalah:

$$wij = \lambda wsim(si, sj) + (1 - \lambda)wdis(si, sj) \quad (1)$$

Dimana ***wsim (si, sj)*** adalah kesamaan cosinus antara vektor dari dua kalimat dan wdis ***(si, sj)*** adalah bobot kalimat itu sedangkan$\boldsymbol{\lambda}\in[0,1]$ adalah parameter yang menyeimbangkan bobot kesamaan, dan jika sebuah kalimat menghubungkan ke dirinya sendiri ***Wii*** maka nilainya 0 (nol). Sedangkan rumus *Wdis*dapat dilihat pada persamaan (2) sebagai berikut :

$$Wdis(sj, si) = \{_{0}^{-1} \quad (2)$$

Nilai -1 adalah nilai dengan discourse connectors (DC) seperti kata “because”, “after”,”before” (Shuzhi Sam Ge dkk 2013)

Tabel 1. Discourse Connector

Hubungan	PENANDA WACANA
Cause-effect	Karena, jadi, karena alasan itulah, karenanya
Temporal	Sebelum, setelah, baru-baru ini, sekarang,
Comparison	nanti, masa lalu
Comparison	Meskipun, tapi, sementara, bagaimanapun, sebaliknya
Expansion	Selain itu, misalnya

Metode *Cosine Similarity* merupakan metode yang digunakan untuk menghitung *similarity* (tingkat kesamaan) antar dua buah kalimat. Metode *cosine similarity* ini menghitung *similarity* antara dua buah kalimat (misalkan D1 dan D2) yang dinyatakan dalam dua buah *vector* dengan menggunakan *keywords* (kata kunci) dari sebuah dokumen sebagai ukuran.

$$CosSim(\mathrm{d}_{i,}, \mathrm{q}_i) = \frac{q_i.d_i}{|\mathrm{q}_i||\mathrm{d}_i|} = \frac{\sum_{j=1}^{t}(\mathrm{q}_{ij}.\mathrm{d}_{ij})}{\sqrt{\sum_{j=1}^{t}(q_{ij})^2.\sum_{j=1}^{t}(d_{ij})^2}} \quad (3)$$

Keterangan :

qij = bobot istilah *j* pada dokumen *i=idf*

dij = bobot istilah *j* pada dokumen *i=idfj*

Pemodelan permeringkatan antar kalimat dalam paper ini digunakan persamaan berikut ini :

$$r(\mathrm{u}_i) = d\sum_{j=1}^{n} r(\mathrm{u}_j)\,\tilde{\mathrm{w}}_{ji} + (1-\mathrm{d}) \quad (4)$$

Dimana $r(\mathrm{u}_i)$ dan u_j dua vertex pada graph dan d adalah parameter antara 0 dan 1.

Non-Negative Matrix Factorization (NMF) dan *Symmetric Matrix Factorization* juga digunakan untuk proses perangkuman dokumen. Dari berbagai kalimat dibagi menjadi beberapa group dan setiap group akan dilakukan ektraksi kalimat pentingnya. Dalam penggunaan metode cluster kalimat bentuk sematiknya akan menemukan dengan cara menelusuran kumpulan sub topik yang tersebunyi (latent) dari dokumen yang secara tidak langsung memberikan informasi tambahan untuk cluster tersebut.

Metode yang digunakan dalam meringkas single dokumen adalah tidak adanya informasi tambahan didalam hasil peringkasan dokumen, metode yang diusulkan adalah memanfaatkan keterkaitan antar kalimat *(sentence mutual effects*) dan mengcluster kalimat berdasarkan hubungan antar kalimat tersbut. Rangking yang kalimat sangat tinggi dan kalimat yang berkaitan dalam satu cluster otomatis akan dipilih sebagai kalimat dalam menyusun rangkuman. Pada penelitian ini memiliki beberapa kontribusi diantaranya adalah:

1. Dalam mengkombinasikan perangkingan kalimat dan clustering menggunakan algoritma perangkuman dokumen berdasarkan nilai iterasi.
2. Sebuah metode clustering untuk kalimat berdasarkan SNMF
3. penggunaan graph berbobot (weighted graph) yang mempertimbangkan rangkaing dari kalimat dalam dokumen beserta hubungan antar kalimat pada cluster.

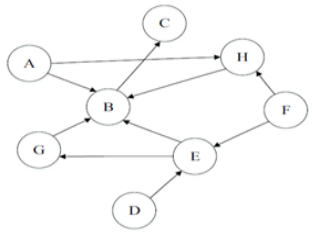
Cara mendapatkan hasil evaluasi pada paper ini memanfaatkan perhitungan ROUGE dan diterapkan pada dataset yang berasal dari 103 tugas akhir mahasiswa yang terdiri dari bagian abstrak dan isi tugas akhir.

B. TextRank Dan Lexrank Pada Dokumen Tunggal

Metode inidiperkenalkanoleh ailin li, tao jiang, qingshuai wang, hongzhi yu, pada journal ieee 20162. Automatic summarization adalah mengekstrak informasi yang paling penting dan berguna secara otomatis menggunakan software dari dokumen asli. Automatic summarization merupan topik penting dibidang natural language processing (pengolahan bahasa alami), machine learning, kecerdasan buatan dan bidang penelitian terkait lainnya.

Pada tulisan ini menjelaskan metode ektraktive yang akan digunakan. Ektraktive summarization terdiri dari kalimat yang di ekstraksi dari dokumen aslinya. Abstrak dapat meringkas isi pokok dari dokumen dan abstrak harus mengandung cukup kata kunci (kata tema) oleh karena itu kata kunci sebagai titik awal penelitian paper ini untuk peringkasan otomatis di tibet yaitu gabungan dari textrank dan lextrank pada dokumen tunggal. Penelitian summarization dalam bahasa cina dan inggris telah banyak mencapai hasil yang bagus tetapi metode yang sudah ada tidak digunakan secara langsung di text tibet karena perbedaan utama pada struktur sintak bahasa.

Di dalam bahasa china terdiri dari subjek-predikat-objek sedangkan dalam bahasa tibet terdiri dari subjek-objek-predikat jadi menurut struktur teks bahasa tibet makalah ini mengusulkan gabungan metode textrank dan lextrank text summarization secara otomatis pada dokumen tunggal. Prinsip dasar algoritma sorting berdasarkan graph adalah ”memilih” atau “rekomendasi” seperti pada gambar. Bila ada hubungan antara titik A dan B di graph dimana titik A menunjuk ke titi B atau titik A merekomendasikan ke titik B dengan asumsi bahwa titik B lebih banyak yang menunjuk maka bisa dikatakan B lebih penting. Selanjutnya pentingnya memilih titik A menentukan pentingnya titik B karena didasarkan fraksi titik yang diperoleh oleh karena itu dijadikan dasar untuk menentukan point yang akan diberikan ke titik B.



Gambar. 2. model sorting graph

Dalam menggunakan algoritma textrank untuk mendapatkan bobot graph ber-arah G={V, E} dimana V adalah kumpulan dari node (vertex) yang mewakili setiap kata dalam dokumen, E adalah tepi (edge) yang mewakili relevansi antar kata-kata. Bobot E*ij* antara simpul Vi dan Vj adalah Wij, dan Wij mewakili kesamaan antara simpul Vi dan Vj diperoleh dengan menghitung jarak kosinus vektor TF-IDF kata-kata yang sesuai. Bobot setiap kata S(v) dapat dihitung dengan koneksi pada simpul S(v) yang lain perhitungan S(V) pada persamaan

$$S(v) = d + (1-d)\sum_{u\in B(v)} \frac{S(u)}{\sum_{j\in F(u)} W_{uj}} \quad (5)$$

B(v) adalah simpul yang menunjuk ke simpul v, F(v) adalah himpunan simpul yang ditunjuk oleh simpul v, d adalah dumping faktor yang berkisar antara 0 sampai 1. Persamaan tersebut mendefinisikan algoritma rekursif untuk menghitung fraksi setiap node dalam graph. Nilai awal S(v) sesuai dengan rekursif graph sampai mencapai nilai konfergen. Pertimbangannya adalah nilai awal dari simpul tidak dapat

.

https://doi.org/10.30743/infotekjar.v5i1.2473

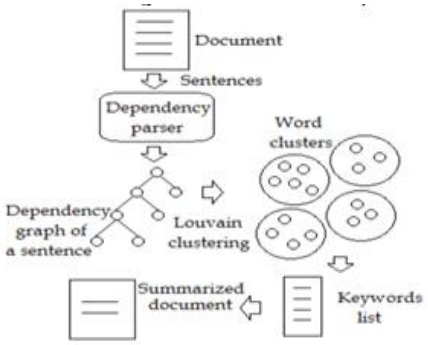
mempengaruhi skor akhir dan hanya mempengaruhi konvergensi algoritma. Setelah konvergensi dihitung maka stabilitas setiap node dalam graph mewakili pusat kata yang sesuai dengan simpul yaitu kemampuan untuk mengetahui ciri dokumen.

Untuk algoritma LexRank, dibutuhkan perhitungan hubungan antara kalimat dan kalimat. Metode LexRank menggunakan istilah frekuensi untukmengukur kesamaan antara kalimatMaka digunakan graph yang tidak berarah G = (S, E). Di antara keduanya, $s \in S$ dinyatakan sebagai kalimat, Sisi (si, sj) E mencerminkan hubungan dari masing-masing kalimat, derajat d dari simpul s adalah jumlah sisi yang terhubung ke s, yang mencerminkan pentingnya informasi yang terkandung dalam kalimat yang sesuai. Semakin besar D, maka semakin banyak jumlah kalimat yang terkait dengan kalimat yang sesuai, semakin banyak pula informasi penting yang terkandung dalam kalimat ini, dan dan sebaliknya. Di sisi lain, jika tingkat simpulnya adalah relatif besar, maka kalimat yang terkait pun juga lebih penting. Dengan cara ini, pertama, paper ini disusun graph G yang tidak berarah dengan menghitung kesamaan antara kalimat; Kedua, kita menggunakan metode perhitungan iterasi untuk menghitung jumlah informasi yang terkandung dalam kalimat sesuai dengan hubungan antara kalimat terakhir pilih satu set kalimat yang paling banyak mengandung informasi sebagai abstrak

Cluster Berbentuk Node-Node Graf

Metode ini diperkenalkan oleh Anyman El-Kilany, Iman Saleh, Journal IEEE 20123. Ringkasan ektraktif dibangun dengan cara pemilihan kalimat penting didalam dokumen. Kalimat penting tersebut ditentukan oleh fitur pada dokumen diantaranya frekuensi kata atau frasa. Kata kunci tersebut dapat menentukan kalimat mana yang harus di ekstraksi dan dapat menentukan posisi kalimat didalam teks. Dalam paper ini kami mengusulkan sebuah metode baru untuk mengekstraksi Single dokumen Summarization. Metode kami menggunakan ketergantungan Grafik kalimat disamping cluster Louvain Algoritma untuk mengekstrak kata kunci dalam dokumen.

Pengumpulan kata dalam cluster adalah bertujuan untuk mengubah kalimat dalam sebuah dokumen menjadi kata-kata. Setiap cluster merepresentasikan sekelompok kata-kata yang koheren(yang berhubungan). Algoritma cluster yang digunakan adalah algoritma lovain. Dan menggunakan dependency graph dalam menentukan hubangan hirarki antar kalimat yang satu dengan kalimat yang lain.



Gambar. 3. Ilustrasi Summarization

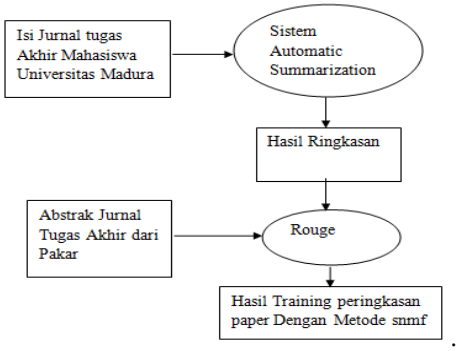
Dalam mencari daftar dari kata kunci kami mengektraksi dari kumpulan data yang dihasilkan atau kumpulan data yang ada. Pada langkah sebelumnya Kata kunci adalah kata yang ditetapkanSkor tertinggi dalam setiap cluster. Skor ini dihitung Menggunakan persamaan 6. Dependency graph diberikan skor 1 kemudian scroe tiap kata adalah persentase skor di setiap induknya. Semakin banyak dependency kata yang ada didalam dependency graph maka dianggap lebih penting karena score kata tersebut didasarkan pada jumlah turunannya atau anaknya dalam dependency graph. Kemudian score ini dinormalisasi oleh jumlah anak dari saudara kandungnya untuk mengurangi tingkat kepentingannya jika saudara yang lain memiliki banyak anak dari pada kata itu sendiri. Skor juga menurun saat kita menurunkan dependency graphnya karena simpul tidak memiliki tanggungan lagi tidak seperti akar dimana semua kata tergantung pada kalimatnya. Untuk menghitung nilai kata S(W) dapat dilihat pada persamaan (6) yang terbetuk berikut ini :

$$S(W) = \frac{CC(W)}{1+CC(W)+\sum_{i=1}^{n(W)} CC(B_i)} * S\big(p(W)\big) \qquad (6)$$

Dimana S(W) adalah nilai kata W, CC(W) adalah jumlah anak dari kata W dalam dependency graph, p(W) adalah induk dari kata tersebut W dalam dependency graph, n(W) adalah jumlah total saudara kata W, Bi adalah satu saudara dari W dalam dependency graph.

C. Blok Diagram Sistem

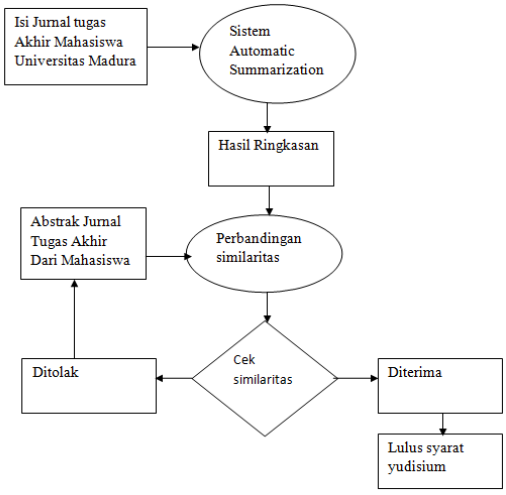
Blok diagram pada Gambar 4 menunjukkan posisi pemanfaatan dari program. Di bawah ini adalah blok diagram dalam menentukan nilai trining peringkasan paper dengan metode snmf yang dilakukan oleh pakar.



Gambar. 4. Blok Diagram Tahap Trining

Blok diagram diatas menjelaskan tentang alur sistem yang dilalui untuk menentukan threshold data latih oleh pakar. Pertama isi dari paper tugas akhir mahasiswa dimasukkan ke sistem summarization kemudian menghasilkan ringkasan, dari hasil ringkasan dimasukkan ke rouge. kemudian dibandingkan dengan abstrak paper tugas akhir mahasiswa dan akan menghasilkan nilai training untuk peringkasan paper dengan metode snmf, hasil dari nilai training masing-masing dijumlahkan semua dan -diambil rata-ratanya, dari rata-rata nilai training ini dibuat sebagai dasar dalam menentukan diterima atau ditolaknya kesesuaian paper tugas akhir mahasiswa universitas madura.

Blok diagram berikut adalah alur yang harus dilalui mahasiswa dalam melakukan checker terhadap seberapa besar tingkat kemiripan abstrak dan isi dokumen paper tugas akhir mahasiswa tersebut.

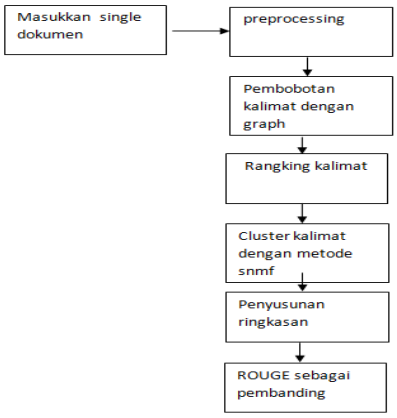


Gambar 5. Diagram Penentuan Data Testing

Dari gambar diatas menjelaskan tentang tahapan testing yang dilakukan oleh mahasiswa Universitas Madura. Tahap pertama mahasiswa memasukkan isi paper tugas akhirnya kemudian sistem secara otomatis akan melakukan peringkasan terhadap isi paper tersebut. Tahap berikutnya adalah memasukkan abstrak untuk dibandingkan dengan sistem kemudian di cek apakah abstrak tersebut lebih dari atau sama dengan nilai rata-rata threshold yang telah ditentukan oleh para pakar pada tahapan data latih jika iya maka hasil dari abstrak dan isi yang diinputkan mahasiswa diterima dan dinyatakan lulus salah satu persyaratan yudisium. Jika tidak maka mahasiswa tersebut direkomendasikan untuk membuat ulang abstrak papernya.

Penggunaan metode yang akan diproses pada sistem untuk pemeriksaan kesesuaian abstrak terhadap isi paper tugas akhir mahasiswa Universitas Madura program studi teknik informatika langkah-langkahnya dapat diurutkan sebagai berikut:

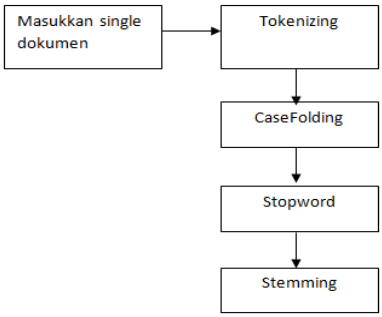
- Tahap awal yaitu Preprocessing
- Membobotkan kalimat hasil preprocessing
- Meraangking kalimat
- Meng-Cluster Kalimat
- Pemilihan Kalimat Representatif
- Penyusunan Ringkasan
- Membandingkan hasil ringkasan kalimat dengan abstrak



Gambar. 6. Blok Diagram proses sistem pemeriksaan kesesuaian abstrak terhadap isi paper tugas akhir mahasiswa Universitas Madura

Penjelasan lengkap blok diagram di atas dapat dilihat pada penjelasan berikut:

1. Preprocessing



Gambar. 7.Diagram Preprocessing

Pada tahapan preprocessing memasukkan isi dari dokumen tunggal berupa isi paper mahasiswa yang akan digunakan sebagai data latih kemudian kalimat atau kata didalam dokumen diproses dan dihitung pada tahap awalnya meliputi Tahapan *tokenizing* adalah tahap mebaca isi dari inputan berupa kata berupa teks, proses awal yang dilakukan adalah membaca isi dokumen secara keseluruhan yang berupa teks. Selanjutnya adalah tahapan case folding tahapan merubah huruf Uppercase menjadi lowercase dengan tujuan menghitung kata yang sama di database, kemudian tahapan stop word yaitu menghilangkan tanda baca kecuali tanda titik (sebagai pembatas akhir dari kalimat) dan tidak dihilangkan. Stop words adalah kumpulan kata umum yang digunakan dalam sistem temu kembali informasi yang tingkat kemunculannya dalam jumlah besar akan diabaikan karena dianggap tidak memiliki makna khusus supaya tidak ikut dihitung ketika perhitungan bobot dari sebuah kalimat atau kata, contoh stop words dalam bahasa indonesia: “atau”, “kami”, “merupakan”, dan lain-lain. Tahapan selanjutnya adalah tahapan Stemming adalah pemotongan kata dengan tujuan mencari kata dasar dari kata yang ada dalam dokumen dengan

https://doi.org/10.30743/infotekjar.v5i1.2473

menggunakan stemmer dalam hal ini kami menggunakan stemmer "sastrawi" dimana stemmer ini dikembangkan untuk dokumen berbahasa Indonesia pada khususnya

2. Pembobot kalimat dengan pemodelan graph

Untuk mendapatkan ringkasan dari sebuah dokumen maka dapat dilakukan dengan cara membuang link yang memiliki bobot graf rendah. Bobot graf yang rendah dapat dipastikan kalimat tersebut tidak banyak memiliki korelasi atau hubungan terhadap kalimat yang lain sehingga tidak masuk sebagai kalimat hasil cluster atau hasil kalimat ringkasan pada sebuah dokumen. Cara memotongnya adalah dengan membatasi nilai-nilai tertentu yang dianggap kecil pada implementasi sistem yang dibuat sebagai sistem peringkasan dokumen. Apabila hasil dari cosinus similarity memiliki nilai 0 (nol) berarti kalimat tersebut dianggap tidak memiliki hubungan antar kalimat. Kalimat-kalimat yang memiliki nilai nol dapat dipastikan tidak dimasukkan dalam peringkasan dokumen tunggal. Hasil dari perhitungan cosinus similarity kemudian dibentuk model graph. Pada model graph dikatakan satu kalimat apabila dipisahkan oleh titik (.) atau koma (,). Tujuan dari pemodelan graph perkalimat ini hanya untuk memudahkan meng identifikasi keterhubungan kalimat yang satu dengan yang lainnya.

3. Rangking kalimat

Tahapan perangkingan kalimat ini adalah kalimat yang sudah terbobotkan yang didapat pada model graph pada proses sebelumnya kemudian di rangking menggunakan algoritma pagerank yaitu melakukakan iterasi dengan batasan tertentu terhadap seluruh kalimat. Dari hasil iterasi tersebut memiliki tujuan untuk menemukan kalimat yang paling banyak memiliki hubungan atau keterkaitan dengan kalimat yang lainnya.

4. Pengelompokan kalimat dengan menggunakan SNMF

Dari hasil kalimat yang sudah melalui tahapan ranking kalimat pada tahapan berikutnya adalah meng-cluster atau mengelompokkan kalimat-kalimat tersebut menjadi beberapa kelompok yang direpresentasikan dengan pengolahan matriks A dengan dimensi nxm menjadi 2 matriks lain yang tidak mengandung nilai negatif, yaitu W yang berdimensi n x r dan H yang berdimensi r x m, dimana r adalah faktor pengalinya dan nilai r < min(n,m). Sedangkan n adalah jumlah kata dan m adalah jumlah

kalimat dalam dokumen.

$$A \approx W.H$$

Matrik A adalah matrik asal non negative sedangkan W adalah matrik fitur semantik non-negative dan matrik variable semantik non negative H masing-masing menggunakan persamaan berikut :

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{(W^T A)_{\alpha\mu}}{(W^T WH)_{\alpha\mu}} \quad \text{.....(4.1)}$$

$$W_{i\alpha} \leftarrow W_{i\alpha} \frac{(AH^T)_{i\alpha}}{(WHH^T)_{i\alpha}} \quad \text{.....(4.2)}$$

Persamaan tersebut digunakan untuk mutiplicative update role untuk matrik H baru dan mutiplicative update role matrik W baru.Untuk mencapai kondisi similar antara *A* dengan *W*H* tersebut, diperlukan suatu kriteria yang dapat dikatakan sebagai *Cost Function*. Pada tahapan berikutnya adalah penggunaan metode SNMF sebagai penyempurnaan dari Algoritma NMF. Algoritma SNMF dapat dijabarkan sebagai berikut :

```
1.  Algoritma Sparse Nonnegative Matrix
    Factorization
2.  Set the sparseness of W and H and the
    target dimension of W and H.
3.  μW=μH=1
4.  while true do
5.  if sparseness constraints on W apply
    then
6.  while true do
7.  W = W- μW (WH – D)
8.  Project each column of W to be non-
    negative, have uncanged L2 norm,
9.  But L1 norm set to achieve desired
    sparseness
10. If E (W,H) decreases then
11. Break
12. End if
13. μW = μW/2
14. if μW < threshold then
15. return
16. end if
17. end while
18. else
19. End if
20. If sparseness constraints on H apply
    then
21. While true do
22. Project each row of H to be non-
    negative,
23. Have unit L2 norm, and L1 norm set to
    achieve desired sparseness
24. If E(W,H) decreases then
25. Break
26. End if
27. μH= μH/2
28. if μH < threshold then
29. return
30. end if
31. end while
32. else
33. End if
34. End while
```

RESULTS AND DISCUSSION

Dengan mengujicobakan terhadap 86 dokumen uji paper tugas akhir mahasiswa Teknik Informatika Universitas Madura kemudian dengan menggunakan dua kombinasi parameter yaitu lambda dan kombinasi parameter dumping factor, nilai parameter yang di ujicobakan untuk lambda yaitu (0.1, 0.3, 0.5) sedangkan nilai dumping factor yaitu (0.2, 0.5, 0.8), kolom pertama adalah urutan dokumen yang diujicobakan kolom ke dua dan seterusnya adalah berisi urutan parameter yang telah diujicoba oleh tiga orang pakar, dalam hal ini pakar yang dilibatkan adalah 3 orang dosen Bahasa Indonesia Universitas Madura, maka dihasilkan pengujian parameter berikut:

Tabel 2. Laporan Hasil Ujicoba dari 3 Orang Pakar

Parameter									
Dok ke	(0.1), (0.2)	(0.1), (0.5)	(0.1), (0.8)	(0.3), (0.2)	(0.3), (0.5)	(0.3), (0.8)	(0.5), (0.2)	(0.5), (0.5)	(0.5), (0.8)
dok1	0.51	0.47	0.56	0.39	0.53	0.46	0.41	0.52	0.52
	0.49	0.38	0.51	0.58	0.48	0.47	0.45	0.55	0.56
	0.44	0.37	0.48	0.35	0.47	0.38	0.37	0.47	0.41
dok2	0.59	0.52	0.60	0.45	0.57	0.62	0.52	0.59	0.52
	0.50	0.39	0.55	0.50	0.58	0.61	0.51	0.54	0.47
	0.53	0.44	0.61	0.46	0.59	0.67	0.53	0.59	0.53
dok3	0.42	0.47	0.42	0.48	0.54	0.53	0.58	0.53	0.53
	0.43	0.46	0.45	0.55	0.56	0.54	0.58	0.53	0.52
	0.33	0.37	0.37	0.39	0.46	0.41	0.47	0.43	0.43
dok4	0.25	0.35	0.27	0.47	0.36	0.35	0.45	0.45	0.43
	0.26	0.40	0.27	0.49	0.37	0.38	0.50	0.51	0.48
	0.26	0.37	0.26	0.51	0.36	0.35	0.44	0.49	0.46
dok5	0.30	0.37	0.32	0.44	0.45	0.42	0.40	0.40	0.38
	0.41	0.39	0.38	0.62	0.58	0.60	0.59	0.54	0.52
	0.47	0.40	0.46	0.72	0.60	0.67	0.62	0.66	0.58
dok6	0.44	0.48	0.46	0.59	0.65	0.63	0.58	0.61	0.58
	0.47	0.48	0.47	0.59	0.62	0.62	0.57	0.61	0.57
	0.48	0.51	0.48	0.60	0.67	0.67	0.62	0.62	0.55
dok7	0.30	0.40	0.37	0.48	0.43	0.48	0.44	0.46	0.43
	0.35	0.40	0.40	0.55	0.51	0.60	0.56	0.55	0.47
	0.36	0.42	0.40	0.55	0.57	0.51	0.51	0.49	0.46
dok8	0.37	0.26	0.31	0.30	0.32	0.29	0.32	0.37	0.32
	0.34	0.32	0.28	0.38	0.43	0.32	0.42	0.47	0.41
	0.35	0.32	0.31	0.41	0.43	0.30	0.43	0.44	0.39
dok9	0.31	0.31	0.26	0.35	0.37	0.29	0.38	0.38	0.34
	0.37	0.31	0.30	0.38	0.42	0.43	0.41	0.39	0.41
	0.37	0.32	0.28	0.41	0.49	0.48	0.50	0.45	0.44
Dok10	0.43	0.45	0.41	0.41	0.48	0.40	0.48	0.51	0.48
	0.557	0.54	0.48	0.45	0.62	0.50	0.50	0.61	0.56
	0.56	0.54	0.47	0.44	0.62	0.50	0.50	0.60	0.57

Dengan penelitian yang telah dilakukan maka di dapatkan rata-rata recall 0.54627 atau dari hasil seluruh dokumen yang dilatih dengan parameter terbaik lamda 0.3 dan dumping factor 0.8. Nilai tersebut menjadi tolak ukur sebagai dasar ditolak atau diterimanya abstrak tugas tugas akhir mahasiswa Universitas Madura.

Pada tahapan berikutnya adalah mencari nilai maksimum recall dengan membandingkan hasil ringkasan sistem dan abstrak yang telah dibuat oleh masing-masing pakar maka didapatkan nilai maksimum recall seperti pada tabel 5.2 dibawah. Dalam menentukan akurasinya, Setelah dirata-rata menghasilkan paremeter tertinggi berada pada kombinasi lambda sama dengan 0.3 dan dumping factor sama dengan 0.8 dengan nilai recall 0.54627. Nilai 0.54627 adalah nilai terbaik setelah diuji cobakan terhadap 86 dokumen training, cara menghitungnya adalah dengan menggunakan nilai maksimum dari masing-masing orang pakar yang membuat abstrak paper tugas akhir seperti pada tabel 4.2 dibawah.

Tabel 3. Hasil Ujicoba 86 Dokumen Nilai Maksimum dari Pakar

Parameter									
No	(0.1), (0.2)	(0.1), (0.5)	(0.1), (0.8)	(0.3), (0.2)	(0.3), (0.5)	(0.3), (0.8)	(0.5), (0.2)	(0.5), (0.5)	(0.5), (0.8)
1	0.51	0.47	0.56	0.58	0.53	0.47	0.45	0.55	0.56
2	0.59	0.52	0.61	0.50	0.59	0.67	0.53	0.59	0.53
3	0.43	0.47	0.45	0.55	0.56	0.54	0.58	0.53	0.53
4	0.26	0.40	0.27	0.51	0.37	0.38	0.50	0.513	0.48
5	0.47	0.405	0.46	0.72	0.60	0.67	0.62	0.66	0.58
6	0.48	0.51	0.48	0.60	0.67	0.67	0.62	0.62	0.58
7	0.36	0.42	0.40	0.55	0.57	0.60	0.56	0.55	0.47
8	0.37	0.32	0.31	0.41	0.43	0.32	0.43	0.47	0.41
9	0.37	0.32	0.30	0.41	0.49	0.48	0.50	0.45	0.44
10	0.56	0.54	0.48	0.45	0.62	0.50	0.50	0.61	0.57
11	0.49	0.50	0.47	0.39	0.60	0.51	0.56	0.54	0.45
12	0.65	0.61	0.59	0.54	0.53	0.49	0.51	0.65	0.60
13	0.44	0.54	0.49	0.51	0.44	0.44	0.52	0.49	0.40
14	0.29	0.41	0.35	0.42	0.43	0.42	0.43	0.56	0.49
Rata-rata	0.47	0.47	0.46	0.53	0.51	0.54	0.53	0.53	0.51

Pada tabel menjelaskan tentang tambahan parameter yang digunakan yaitu dumping factor 0.1, 0.3, 0,5 dan lambda 0.85 serta penambahan nilai recall, precision dan f-score hasilnya adalah sebagai berikut:

Tabel 4. Penambahan nilai recall, precision dan f-score

Dok ke	**Recall (0.1), (0.85)**	**Precision (0.1), (0.85)**	**F-Score (0.1), (0.85)**	**Recall (0.3), (0.85)**	**Precision (0.3), (0.85)**	**F-Score (0.3), (0.85)**	**Recall (0.5), (0.85)**	**Precision (0.5), (0.85)**	**F-Score (0.5), (0.85)**
1	0.55	0.08	0.14	0.49	0.07	0.12	0.50	0.07	0.12
	0.54	0.06	0.11	0.47	0.05	0.10	0.50	0.06	0.10
	0.44	0.12	0.19	0.39	0.10	0.16	0.42	0.11	0.18
2	0.55	0.05	0.09	0.50	0.04	0.08	0.52	0.06	0.11
	0.48	0.04	0.08	0.54	0.05	0.09	0.52	0.06	0.12
	0.54	0.05	0.10	0.51	0.05	0.10	0.56	0.07	0.13
3	0.45	0.07	0.13	0.46	0.08	0.14	0.59	0.09	0.16
	0.48	0.09	0.15	0.50	0.09	0.16	0.57	0.10	0.17
	0.39	0.08	0.14	0.39	0.09	0.15	0.49	0.10	0.17
4	0.39	0.05	0.09	0.41	0.06	0.11	0.52	0.06	0.12
	0.43	0.05	0.09	0.39	0.05	0.09	0.40	0.04	0.08
	0.39	0.05	0.09	0.40	0.05	0.10	0.44	0.06	0.11
5	0.38	0.06	0.11	0.43	0.07	0.13	0.49	0.10	0.17
	0.45	0.06	0.10	0.52	0.07	0.12	0.64	0.11	0.19
	0.45	0.04	0.08	0.59	0.06	0.11	0.59	0.08	0.14

https://doi.org/10.30743/infotekjar.v5i1.2473

6	0.30	0.04	0.08	0.30	0.06	0.10	0.31	0.04	0.07
	0.37	0.04	0.08	0.39	0.06	0.10	0.38	0.04	0.07
	0.38	0.03	0.06	0.41	0.04	0.08	0.42	0.03	0.06
7	0.40	0.05	0.10	0.50	0.09	0.15	0.45	0.10	0.16
	0.41	0.04	0.08	0.54	0.07	0.13	0.51	0.08	0.14
	0.42	0.04	0.07	0.59	0.06	0.12	0.56	0.07	0.13
8	0.33	0.05	0.10	0.37	0.03	0.07	0.36	0.05	0.09
	0.30	0.03	0.06	0.39	0.02	0.05	0.41	0.03	0.06
	0.31	0.03	0.06	0.34	0.02	0.04	0.42	0.03	0.06
9	0.26	0.03	0.06	0.33	0.06	0.11	0.29	0.05	0.08
	0.28	0.03	0.06	0.41	0.09	0.15	0.35	0.06	0.11
	0.24	0.04	0.07	0.48	0.11	0.19	0.35	0.07	0.12
10	0.44	0.05	0.09	0.52	0.07	0.12	0.55	0.05	0.09
	0.44	0.04	0.08	0.61	0.07	0.12	0.61	0.05	0.09
	0.44	0.04	0.07	0.59	0.06	0.11	0.62	0.04	0.08

Dari ujicoba abstrak mahasiswa terhadap isi paper yang diinputkan, maka didapatkan hasil sepuluh diterima oleh sistem dan delapan belas ditolak oleh sistem karena tidak mencapai parameter yang telah ditentukan yaitu recall 0.54627. Sedangkan dalam pengujian perbandingan paper menurut sistem dibandingkan pengujian manual yang dilakukan oleh pakar didapatkan hasil tebakan yaitu 7 tebakan salah dan 21 tebakan benarsehingga jika di persentase tingkat akurasi mencapai$\frac{21}{28} x100\% = 75\%$ dan tingakat kesalahan tebakan mencapai $\frac{7}{28} x100\% = 25\%$. Dapat dilihat pada tabel berikut :

Tabel 5. Hasil Diterima dan Ditolak Paper Mahasiswa

Dokumen	Status	Recall
1	DITERIMA	0.56866
2	DITERIMA	0.60069
3	DITERIMA	0.56809
4	DITERIMA	0.56028
5	DITOLAK	0.35492
6	DITOLAK	0.40988
7	DITOLAK	0.42678
8	DITOLAK	0.50512
9	DITOLAK	0.43059
10	DITOLAK	0.34257
11	DITOLAK	0.48054
12	DITOLAK	0.33994
13	DITOLAK	0.50556
14	DITERIMA	0.58999
15	DITERIMA	0.55593
16	DITOLAK	0.36892
17	DITOLAK	0.42547

CONCLUSIONS

Seluruh kesimpulan yang didapat dari penelitian ini adalah sebagai berikut:

Dengan melibatkan 3 orang pakar dalam pembuatan abstrak paper tugas akhir mahasiswa Universitas Madura dan 2 kombinasi parameter lamda serta dumping factor dalam menentukan paremeter terbaik maka diperoleh recall sebesar 0.54627 sebagai tolak ukur diterima atau ditolaknya paper mahasiswa oleh sistem.

Dengan menggunakan metode SNMF (Sparse Non Negative Matrix Factorization) dalam cluster kalimat pada dokumen tunggal maka dihasilkan ringkasan dokumen dengan mengambil 2 kalimat tertinggi pada masing-masing cluster kalimat.

Dalam penentuan nilai recall yang telah dilatihkan terhadap dokumen abstrak tugas akhir sebanyak 86 dokumen latih dan 17 dokumen yang diujikan sehingga menghasilkan sebanyak 6 dokumen diterima oleh sistem dan 11 dokumen ditolak oleh sistem.

ACKNOWLEDGMENT

Terimakasih kepada bapak reddy alexandro H, dan bapak Hendrawan Armanto dalam memberikan masukan terhadap penulisan paper ini hingga selesai.

REFERENCES

[1] Shuzhi Sam Ge, Zhengchen Zhang, Hongsheng He (2013), "Weighted graph Model based sentence clustering and Rangking for Document Summarization", *Journal of IEEE,2013*

[2] Anyman El-Kilany,Iman Saleh (2012), "Unsupervised Document Summarization Using Clusters of Dependency Graph Nodes", Journal of *International Conference on Intelligent Systems Design and Applications (ISDA),* IEEE (2012),hal557-561.

[3] Ailin Li, Tao Jiang, Qingshuai Wang, Hongzhi Yu (2016), "The Mixture of TextRank and LexRank Techniques of Single Document Automatic Summarization Research in Tiben", *Journal of International Conference on Intelligent Human-Machine Systems and Cybernetics*, IEEE (2016) hal. 514-519.

[4] Patrik O. Hoyer, (2004), "Non-Negative Matrix Factorization with Sparseness Constrains", *International Journal of Machine Learning Research*, 5,(2004), hal. 1457-1459.

[5] Lee, D. D., & Seung, H. S. (2001). Algorithm for non-negative matrix factorization. *Advance in Neural Information Processing Systems, 13* , 556-562.

[6] Sarkar, Kamal, (2013), "*Automatic single document Text Summarization Using Key Consepts in Document*," *The Journal of J Inf Process Syst*, vol. 9,no.4, pp. 602–620, 2013.

[7] Rolly Intan, Andrew Defeng, "**HARD:Subject-based *Search Engine* menggunakanTF-IDF dan Jaccard's Coefficient". Universitas Kristen petra, Surabaya.**

[8] Asian J. (2007) "*Effective Techniques for Indonesian Text Retrieval*". PhD thesis School of Computer Science and Information Technology RMIT University Australia.

[9] Anyman El-Kilany,Iman Saleh (2012), "Unsupervised Document Summarization Using Clusters of Dependency Graph Nodes", Journal of *International Conference on Intelligent Systems Design and Applications (ISDA),* IEEE (2012),hal557-561.

AUTHOR(S)BIOGRAPHY

Irwan Darmawan, S.Kom., M.Kom.
Adalah salah satu dosen Informatika di universitas Madura Kabupaten Pamekasan. Menyelesaikan S-1 di Universitas Trunojoyo dan S-2 Di STTS Surabaya.