e-ISSN 2338-5197

Text Mining Untuk Mengklasifikasi Judul Skripsi Menggunakan Metode TF IDF Dan Algoritma C4.5

Ansari[a,1], Tedy Setiadi [a,2], Lisna Zahrotun[a,3]

[a] Program Studi Teknik Informatika, Universitas Ahmad Dahlan,
Prof. Dr. Soepomo, S.H., Janturan, Umbulharjo, Yogyakarta 55164
[1]ansari12018131@webmail.uad.ac.id; [2]tedy.setiadi@tif.uad.ac.id; [3]lisna.zahrotun@tif.uad.ac.id

Abstrak

Skripsi merupakan karya ilmiah mahasiswa yang disusun dalam rangka memenuhi sebagian syarat penyelesaian studi pada program strata satu (S-1). Langkah awal yang dilakukan mahasiswa dalam mendaftar skripsi adalah mendaftarkan judul skripsi kepada Koordinator TA. Setelah mahasiswa mendaftar maka langsung di tentukan pembimbing dan pengujinya, tetapi dalam hal ini Koordinator TA menemui kesulitan dalam memilah skripsi setiap mahasiswa sesuai dengan bidang minatnya karena sebagai acuan untuk menentukan pembimbing dan pengujinya. Dengan banyaknya mahasiswa yang mendaftarkan skripsi, maka terdapat data skripsi yang melimpah. Data dari judul skripsi bisa dimanfaatkan untuk pencarian pola klasifikasi. Pola tersebut nantinya bisa di terapkan kedalam program untuk memudahkan Koordinator TA dalam menentukan skripsi mahasiswa sesuai dengan bidang minatnya untuk penentuan pembimbing dan pengujinya.Proses-proses text mining seperti tokenizing, filtering, stemming, cleaning, pembobotan, klasifikasi untuk pembentukan pola, dan pengujian keakurasian. Pembobotan dilakukan dengan TF-IDF sedangkan klasifikasi menggunakan Algoritma C4.5, untuk pengujian keakurasian dilakukan dengan menggunakan confussion matrix. Dalam penelitian yang telah dilakukan membuktikan bahwa TF-IDF dan Algoritma C4.5 dapat diterapkan untuk mengklasifikasikan judul skripsi. Data set sebanyak 142 dengan menggunakan 130 data training dan 12 data testing, tingkat akurasi yang didapat mencapai 92%.

Kata Kunci: *Text Mining*, TF-IDF, Algoritma C4.5, Judul Skripsi.

1. Pendahuluan

Skripsi merupakan karya ilmiah mahasiswa yang disusun dalam rangka memenuhi sebagian syarat penyelesaian studi pada program strata satu (S-1). Karya ilmiah tersebut berupa laporan penelitian, baik penelitian lapangan, penelitian pustaka, penelitian laboratorium, maupun penelitian pengembangan [1] .

Skripsi di Teknik Informatika UAD dilakukan sebagai syarat untuk menyelesaikan studi selama masa perkuliahan program S1. Prodi Teknik Informatika UAD terdapat tiga bidang minat yaitu, bidang minat "*Sistem Informasi*", bidang minat "*Web dan Mobile Computing*", dan bidang minat "*Soft Computing dan Multimedia*". Berdasarkan data semester ganjil tahun 2013/2014, mahasiswa yang mengikuti skripsi cukup banyak yaitu 181 peserta dari semua bidang minat [2].

Langkah awal yang dilakukan mahasiswa dalam mendaftar skripsi adalah mendaftarkan judul skripsi kepada Koordinator TA. Setelah mahasiswa mendaftar maka langsung di tentukan pembimbing dan pengujinya, tetapi dalam hal ini Koordinator TA menemui kesulitan dalam memilah skripsi setiap mahasiswa sesuai dengan bidang minatnya karena sebagai acuan untuk menentukan pembimbing dan pengujinya. Dengan banyaknya mahasiswa yang mendaftarkan skripsi, maka terdapat data skripsi yang melimpah. Data dari judul skripsi bisa dimanfaatkan untuk pencarian pola-pola atau pengetahuan yang tersembunyi dari data tersebut. Pola-pola tersebut nantinya bisa di terapkan kedalam program untuk memudahkan Koordinator TA dalam menentukan skripsi mahasiswa sesuai dengan bidang minatnya untuk penentuan pembimbing dan pengujinya.

e-ISSN 2338-5197

Dengan demikian harapannya ada suatu aplikasi yang dapat membedakan judul skripsi sesuai dengan bidang minatnya. Aplikasi diharapkan dapat memberikan pengetahuan tentang proses pengklasifikasian sesuai dengan bidang minatnya.

2. Kajian Pustaka

Penelitian ini mengacu pada penelitian terdahulu [3]. Pada penelitian tersebut dibahas mengenai analisis kesesuaian bidang minat yang diambil mahasiswa sehingga ilmu yang diperoleh dalam matakuliah bidang minat dapat diterapkan dalam tugas akhir mahasiswa. Dalam penelitian tersebut cukup mempunyai keunggulan akan tetapi tidak menjelaskan tahapan-tahapan *text mining* didalam penerapannya

Penelitian ini juga mengacu pada penelitian [4]. Pada penelitian tersebut membahas mengenai analisis pengklasifikasian dokumen Tugas Akhir Mahasiswa yang belum diketahui kategorinya, setelah dilakukan *preprocessing* didapatlah kata-kata yang berhubungan dengan teknik informatika. Pada penelitian ini sudah menjelaskan sebagian cara kerja *text mining* namun tidak menghasilkan pola sebagai pengetahuan.

Pada penelitian ini akan dikembangkan sebuah program yang menjelaskan tahapan text mining, dapat menghasilkan pola sebagai pengetahuan, dan dapat menampilkan hasil tes pengklasifikasian dari penerapan metode TF IDF dan algoritma C4.5.

1. Text Mining

Text Mining memiliki definisi menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen, dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen [5].

2. TF IDF

TF IDF merupakan salah satu metode pembobotan yang paling popular untuk perangkingan berdasarkan kemunculan kata di dalam suatu teks atau mesin pencari. TF IDF merangking berdasarkan tingkat kemunculan suatu kata dalam suatu teks [6].

Rumus TF IDF adalah sebagai berikut :

IDF = log (d/df)

Keterangan :

IDF : Nilai Invers Document Frecuency
df : Jumlah kemunculan dokumen
d : Jumlah dokumen

3. Algoritma C4.5

Algoritma C4.5 adalah salah satu metode untuk membuat *decision tree* berdasarkan *training* data yang telah disediakan. Diperkenalkan oleh Quinlan [7] yang merupakan pengembangan dari ID3. Beberapa pengembangan yang dilakukan pada C4.5 antara lain bisa mengatasi *missing value*, bisa mengatasi *continue data*, dan *pruning*.

Rumus C4.5 untuk menghitung entropi:

$$Entropy(S) = \sum_{i=1}^{n} -pi * log2pi \quad \text{...1)}$$

Keterangan :
S : Himpunan kasus
A : Fitur

e-ISSN 2338-5197

n : Jumlah partisi S
pi : Proporsi dari Si terhadap S
Rumus C4.5 untuk menghitung entropi dan gain masing-masing atribut:

$$Gain\ (S,A) Entropy(S) - \sum_{i=0}^{n} \frac{|Si|}{|S|} * \text{Entropy}\ (Si) \quad ...2)$$

Keterangan :
S : Himpunan kasus
A : Atribut
n : jumlah partisi atribut A
|Si|: Jumlah kasus pada partisi ke i
|S| : Jumlah kasus dalam S

4. Confussion Matrix

Pengujian keakurasian sistem dilakukan dengan menggunakan metode confussion matrix, suatu metode untuk akurasi yang sudah biasa dalam data mining [8].

Rumus penghitungan *accuracy* dan *error rate*:

a. *Accuracy* = (a + d) / (a + b + c + d).
b. *Error rate* = (b + c) / (a + b + c + d).

Keterangan dari masing-masing atribut :
a = jika hasil prediksi positif dan data sebenarnya positif.
b = jika hasil prediksi positif sedangkan nilai sebenarnya negatif.
c = jika hasil prediksi negatif sedangkan nilai sebenarnya positif.
d = jika hasil prediksi negatif dan nilai sebenarnya negatif.

3. Metode Penelitian

7. Obyek Penelitian

Dalam penelitian ini yang menjadi obyek penelitian adalah mengklasifikasi judul skripsi dengan menggunakan metode pembobotan TF-IDF dan algoritma C4.5. Diharapkan Algoritma C4.5 dapat mengklasifikasikan judul skripsi dengan akurat.

8. Tahapan Text Mining

Pada bagian ini menjelaskan tentang tahapan-tahapan pengklasifikasian judul skripsi, untuk memudahkan dalam penyusunan pembahasan.

a. *Load*/Memasukkan Data Judul Skripsi

Tahapan awal yang dilakukan adalah memasukkan data judul skripsi. Data judul skripsi diambil dari mahasiswa yang mendaftar skripsi pada semester ganjil tahun ajaran 2013/2014.

b. Tokenizing

Pada tahapan ini judul skripsi dipisahkan menjadi perkata. Dalam proses tokenizing juga dilakukan penghapusan tanda baca seperti ‘!’, ‘-’, ‘.’, ‘,’ dan lain-lain.

c. Filtering

Tahap *filtering*, mengambil kata-kata dari hasil *tokenizing* dan membuang kata-kata yang kurang penting. Kata yang dihapus adalah kata hubung atau konjungsi yaitu “yang”, “dan”, “dengan”, “di”, dan lain-lain.

d. Stemming

Tahap *Stemming*, proses setelah *filtering*. Judul skripsi yang masih memiliki imbuhan maka akan dikembalikan ke kata dasar, misalnya “melihat” dikembalikan ke kata dasar menjadi “lihat”.

e. Cleaning

e-ISSN 2338-5197

Data judul skripsi yang tidak mengandung kata kunci maka judul tersebut akan di hapus.

f. Pembobotan TF IDF

Proses pembobotan dari sebuah kata yang dinilai dari kemunculan kata yang cocok dengan kata kunci yang telah tersedia. Pembobotan TF-IDF akan menghasilkan kategori dari masing-masing judul skripsi.

g. Algoritma C4.5 dan Pola

Setelah proses pembobotan dilakukan maka akan diklasifikasi dengan menggunakan algoritma C4.5 yang hasilnya akan menjadi pola dan pengetahuan.

9. Pengujian Keakurasian

Untuk menguji akurasi, metode *Confussion matrix* digunakan untuk mengetahui sebaik apakah Algoritma C4.5 dalam hal akurasi.

4. Hasil Dan Pembahasan

2. Tahapan Text Mining

a. *Load*/Memasukkan Data Judul Skripsi

Jumlah data judul skripsi yang akan diolah yaitu 181 data judul skripsi. Data judul skripsi akan diuji dengan dua kali pengujian, pengujian pertama yaitu 117 data training dan 25 data testing kemudian pengujian kedua yaitu 130 data training dan 12 data testing.

b. Tokenizing

Judul skripsi di pisah menjadi perkata dan penghapusan tanda baca seperti ‘!’, ‘-’, ‘.’, ‘,’ dan lain-lain.

Judul:

Analisis Keamanan Data Pada Sistem **E-Commerce** dengan Kerangka Kerja Zachman
Hasil Tokenizing:
analisis-keamanan-data-pada-sistem ecommerce-dengan-kerangka-kerja-zachman

c. Filtering

Membuang kata-kata yang kurang penting. Kata yang dihapus adalah kata hubung atau konjungsi yaitu “yang”, “dan”, “dengan”, “di”, dan lain-lain.

Judul:

Analisis Keamanan Data **Pada** Sistem E-Commerce **dengan** Kerangka Kerja Zachman
Hasil Filtering:
analisis-keamanan-data-sistem ecommerce-kerangka-kerja-zachman

d. Stemming

Judul skripsi yang masih memiliki imbuhan maka akan dikembalikan ke kata dasar

Judul:

Analisis **Keamanan** Data Pada Sistem E-Commerce dengan **Kerangka** Kerja Zachman
Hasil Stemming:
analisis-aman-data-sistem ecommerce-rangka-kerja-zachman

e. Cleaning

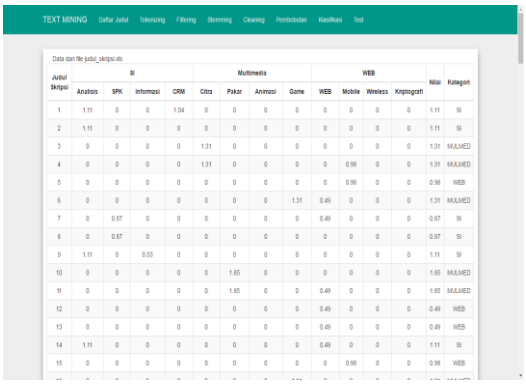
Dari 181 data judul skripsi setelah diproses *cleaning* judul skripsi yang akan diproses ke tahap berikutnya tersisa 142 data judul skripsi.

f. Pembobotan TF IDF

Hasil dari pembobotan pencocokkan kata dengan kata kunci didapat 81 judul skripsi masuk bidang minat SI, 24 judul skripsi masuk bidang minat MULMED, 37 judul skripsi masuk bidang minat WEB.

e-ISSN 2338-5197

Tampilan penghitungan TF IDF pada program, terlihat pada gambar 1:

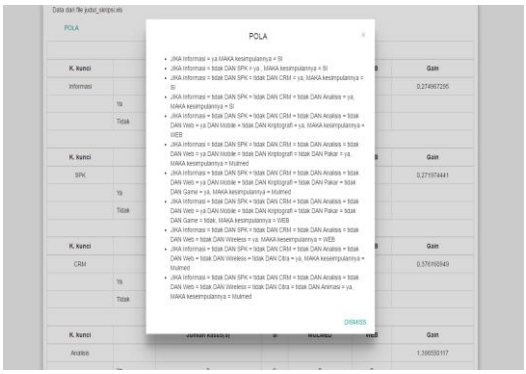


Gambar 1 Tampilan TF IDF

g. Algoritma C4.5 dan Pola

Setelah proses pembobotan dilakukan maka akan diklasifikasi dengan menggunakan algoritma C4.5 yang hasilnya akan menjadi pola dan pengetahuan.

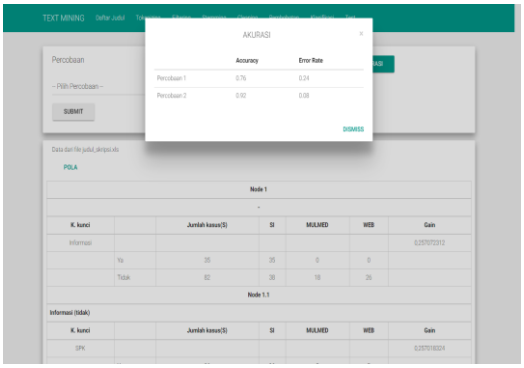
Tampilan pola dari penerepan algoritma C4.5 pada program, terlihat pada gambar 2:



Gambar 2 Tampilan C4.5

3. Pengujian Akurasi

Tampilan pada program, uji akurasi menggunakan metode confussion matrix, terlihat pada gambar 3:



Gambar 3 Uji Akurasi

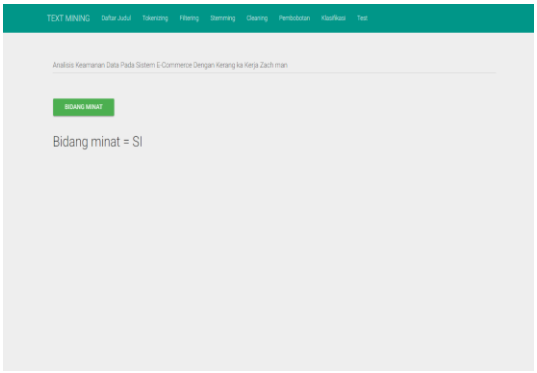
Pengujian yang pertama dengan data training 117 dan data testing 25 didapat accuracy sebesar 76% dan error rate sebesar 24%. Pengujian kedua dengan data training 130 dan data testing 12

e-ISSN 2338-5197

didapat accuracy sebesar 92% dan error rate sebesar 8%. Dapat disimpulkan jika data training lebih banyak dibanding dengan data testing maka nilai akurasinya tinggi sedangkan kebalikannya jika data training lebih sedikit dibandingkan dengan data testing maka nilai akurasinya rendah.

4. Hasil Test

Tampilan test judul skripsi pada program, terlihat pada gambar 4:



Gambar 4 Tampilan Test

Inputan pada test judul skripsi pada gambar 2 mengandung kata kunci "analisis" dan kesimpulannya judul skripsi tersebut masuk bidang minat SI

5. Penutup

5.1. Kesimpulan

a. Diterapkannya metode TF IDF dan algoritma C4.5 ke dalam program yang dapat mengklasifikasi judul skripsi.
b. Terujinya algoritma C4.5 dalam hal keakurasian. Hasil akurasi sistem yang dihasilkan 92%, dengan data training 130 dan data testing 12.

Daftar Pustaka

[1] Laksono, K. (2014). *Pedoman Penulisan Skripsi*. Surabaya: Universitas Negeri Surabaya.
[2] Tif.uad.ac.id. (2013). Peserta Tugas Akhir Ganjil 13-14. *Universitas Ahmad Dahlan*. Retrieved from tif.uad.ac.id
[3] Meliana, N. (2008). Deteksi Kesesuaian Bidang Minat Terhadap Proposal Tugas Akhir Mahasiswa. *UKDW*.
[4] Prayitno, D. (2012). Penerapan Naive Bayes Classifier dalam Pengklasifikasian Jurnal Tugas Akhir. *Universitas Pendidikan Indonesia*, 10–38.
[5] Raymond J. Mooney. (2006). *Machine Learning Text Categorization*. Austin: University of Texas.
[6] Paik, J. (2013). TF IDF Scheme For Efective Ranking and Development Information Retrieval. Dublin.
[7] Quinlan, J. . (1996). Improved Use of Continous Attributes in C4.5. *Artificial Inntelligence Research*, 77–90.
[8] Kao, A. (2007). *Natural Language Processing and Text Mining*. United States of America: Business Media.