JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 6 Nomor. 1, April 2021
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

KLASIFIKASI BIBLIOGRAFI OTOMATIS MENGGUNAKAN C4.5 DAN INFORMATION GAIN

MUHAMMAD NUR AKBAR

Teknik Informatika Fakultas Sains dan Teknologi
Universitas Islam Negeri Alauddin Makassar

E-mail: muhammad.akbar@uin-alauddin.ac.id

ABSTRAK

Permasalahan yang dibahas pada penelitian ini mengenai klasifikasi bibliografi. Klasifikasi dilakukan dengan memproses data-data dari berbagai sumber referensi yang diberikan. Metode yang diterapkan dalam pengklasifikasian adalah C4.5 dengan sebelumnya dilakukan beberapa tahap *preprocessing*. C4.5 yang digunakan untuk proses *text mining* karena memiliki akurasi dan kecepatan yang sangat tinggi dengan algoritma yang sederhana. Digunakan pula *Information Gain* untuk evaluasi atribut yang dipilih dalam mengklasifikasikan dokumen.

***Kata Kunci*: Text mining, C.45, bibliography, feature selection, Information Gain;**

I. PENDAHULUAN

Ketersediaan data saat ini sangat melimpah dan beragam. Dari data yang ada dapat ditarik informasi apabila pola keterhubungan antar data dapat diketahui.

Data mining adalah proses untuk menemukan pola data yang menarik dan berguna untuk mendapatkan informasi dari data-data yang tersedia dengan cara menganalisa data dan menggunakan algoritma-algoritma tertentu untuk memproses data. Dalam penangan kasus data mining dikenal beberapa istilah seperti *noise* dan *missing value*. *Noise* adalah data yang mengandung *error* atau *value* yang tidak wajar yang timbul karena faktor kesalahan *entry* oleh manusia, komputer *error*, atau terdapat kesalahan saat pengiriman data. Sedangkan *missing value* adalah data atau informasi yang hilang atau tidak tersedia akibat faktor *non sampling error.*

Salah satu jenis data yang tersedia banyak adalah data dalam bentuk *text* atau dokumen. Bidang khusus dari *data mining* yang menangani pencarian pola *data text* disebut dengan *text mining*. Tujuan dari *text mining* ini adalah pengambilan informasi berkualitas dari *text* tersebut.

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 6 Nomor. 1, April 2021
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

Dalam *text mining* dikenal empat macam atribut yang sering digunakan yaitu *character, word, term,* dan *concept*.

Pada bibliografi ditampilkan data dari sumber-sumber referensi. Data-data tersebut kemudian diproses untuk menemukan pola yang dapat mengelompokkan setiap referensi kedalam kategori tertentu.

Dalam permasalahan ini disediakan data dari beberapa sumber referensi. Data tersebut memiliki atribut: *bibliography type*, ISBN, *identifier, author, title, publication, volume, number, month, page, year, address, note*, URL, *booktitle, chapter, edition, series, editor, publisher, report, howpublished, institution, organizations, school, annote, custom 1, custom 2, custom 3, custom 4, custom 5,* dan *class*.

Disediakan dua buah kelompok data yang memiliki atribut yang sama, yaitu *data training* dan *data testing*. *Data training* terdiri dari 1200 baris data yang digunakan untuk menemukan pola dokumen yang dikelompokkan menjadi kelas tertentu. Sedangkan *data testing* terdiri dari 593 baris data.

II. METODE PENELITIAN

A. PREPROCESSING

Data training yang diberikan terdiri atas 31 atribut, namun data yang digunakan untuk menyelesaikan masalah ini hanya diambil dari sembilan atribut sedangkan atribut-atribut lainnya diabaikan dengan pertimbangan sebagai berikut:

- Atribut *ReportType* dan *howpublised* bernilai *null* untuk semua baris.
- Atribut *BibliographyType* dan atribut *Year* memiliki jumlah dari setiap jenis *value* yang cenderung sama untuk masing-masing kategori sehingga dianggap tidak dapat mewakili kategorinya. Sebagai contoh dapat dilihat pada data kategori B dan C dibawah ini:

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 6 Nomor. 1, April 2021
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

Tabel 1 .Perbandingan *value* kategori B dan kategori C

Value	Jumlah *value* pada kategori	
	B	C
Article	187	183
Inproceeding	3	4
Technical Report	3	0
Book	5	10
Misc	1	1
Incollection	1	1
Phd Thesis	0	1
Unpublished	0	0

- Atribut *Identifier* memiliki *value* yang berbeda-beda untuk setiap baris sehingga setiap *value* hanya merepresentasikan barisnya namun tidak mewakilkan kategorinya.

- Atribut ISBN, *Volume, Number, Month, Page, Address, Note, Booktitle, Chapter, Edition, Series, Editor, Publisher, Institution, Organizations, School, Annote* memiliki banyak *value* yang *null* seperti yang terlihat pada tabel di bawah, sedangkan yang tidak *null* ditemukan beberapa *noise* dan *value*nya sangat beragam sehingga tidak dapat mewakili kategorinya.

Tabel 2 Persentasi jumlah value pada atribut yang diabaikan

Atribut	*Value* tidak *null*	*null*
ISBN	7,3%	92,7%
Volume	55,7%	44,3%
Number	47,2%	52,8%
Month	19,8%	80,2%
Page	72,9%	27,1%
Address	9,4%	90,6%
Note	0,08%	99,92%
Booktitle	10%	90%
Chapter	2,3%	97,7%
Edition	0,6%	99,4%
Series	0,6%	99,4%
Editor	2,7%	97,3%
publisher	15,6%	84,4%
Institution	1,4%	98,6%
Organization	0,16%	99,84%
School	0,25%	99,75%
Annote	4,4%	95,6%

Dengan pertimbangan diatas, maka dalam mengolah data atribut yang digunakan adalah atribut *Author*, dengan pertimbangan setiap baris memiliki *value*

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 6 Nomor. 1, April 2021
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

dan ditemukan kata-kata yang sama pada buku berbeda. Selain itu, mengingat pada umumnya setiap penulis (*author*) memiliki keahlian pada suatu bidang yang ditekuninya sehingga ia akan menulis tulisan-tulisan dengan tema yang mirip seputar bidang keahliannya. Oleh karena itu, atribut *author* dianggap relevan untuk mewakili kategorinya.

Langkah selanjutnya adalah menyatukan sembilan atribut diatas untuk kemudian diolah dalam proses *text mining* sebagai satu dokumen teks yang kemudian menghasilkan 8932 atribut yang merupakan representasi dari setiap kata yang telah melalui tahap *preprocessing*.

Adapun dalam *preprocessing* ini dilakukan beberapa langkah untuk memproses data berupa *text*, yaitu:

1. *Word Token*

   Word token adalah tahap pemotongan *text* berdasarkan tiap kata yang menyusunnya. Contoh dari tahap ini sebagai berikut:

   Input : *DNA repeats are a common atribut of most genomic sequences.*

   Hasil token : DNA repeats are a common atribut of most genomic sequences

2. *Stopword Removal*

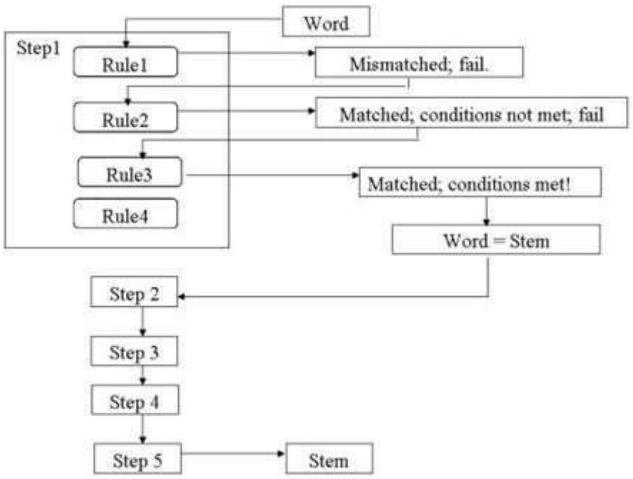
   Tujuan dari tahap *stopword removal* adalah untuk menyaring kata-kata yang tidak layak untuk dijadikan kata kuci sehingga kata-kata tersebut dapat dihapus dari *text*. Beberapa contoh term yang dapat dihapus adalah ‘a’, ’about’, ‘because’, ‘and’, ‘cant’, ‘do’, ‘enough’, ‘ever’, ‘for’, ‘get’, ‘have’, ‘however’, ‘is’, ’mostly’, dll. Proses ini dapat membuat pencarian pola untuk klasifikasi menjadi lebih efektif karena hanya akan menyisakan kata-kata yang dapat menjadi *keyword* tiap kategori. Sebagai contoh, bila hasil

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 6 Nomor. 1, April 2021
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

dari tahap *word token* dilakukan proses stopword removal, maka kata are, a, of, most akan dihapus.

3. *Stemming*

Stemming adalah tahap mentranformasi kata-kata dalam *text* menjadi bentuk kata dasarnya. Algoritma yang digunakan dalam proses *stemming* ini adalah Algoritma *Porter*. Adapun bagan alir algoritma *Porter Stemmer* digambarkan seperti di bawah ini:



Gambar .1 Flow Diagram Algoritma *Porter Stemmer*

Data yang diberikan berupa teks berbahasa Inggris, oleh sebab itu proses *stemming* disesuaikan agar memenuhi aturan bahasa yang benar.

4. *Term Weighting*

Term Weighting atau pembobotan dilakukan terhadap kata-kata yang dianggap dapat mewakili kategorinya. Bobot tersebut merepresentasikan besarnya peran dari masing-masing *term* dalam mewakilkan kategorinya. Dalam permasalahan ini dilakukan pembobotan menggunakan *Tf-Idf*.

Tf-Idf adalah cara untuk memberikan bobot berdasarkan kemunculan suatu *term* dalam dokumen. Pemilihan metode *Tf-Idf* dalam tahap ini dikarenakan nilai precision dan recall yang lebih baik serta waktu yang dibutuhkan untuk eksekusi lebih cepat bila dibandingkan dengan metode lain. Rumus *Tf-Idf*:

$$w_{tf} = tf \times idf$$

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 6 Nomor. 1, April 2021
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

$$w_{tf} = tf \times \log \frac{N}{df}$$

Keterangan:

w_{tf} = bobot term t_f terhadap dokumen d_f

tf = jumlah kemunculan *term* t_f dalam d_f

N = jumlah dokumen yang dibandingkan

df = jumlah dokumen yang mengandung term t_f

5. *Feature Selection*

Feature Selection adalah metode untuk mereduksi atribut yang jumlahnya sangat banyak dengan cara memilih atribut yang paling berpengaruh dalam kegiatan pemodelan data. Metode *atribut selection* yang digunakan untuk menangani masalah ini adalah *information gain* (IG).

Metode *information gain* menggunakan konsep menghitung keteracakan kata. Sebuah kata akan memiliki nilai IG yang tinggi apabila kata tersebut muncul di banyak dokumen dalam suatu kelas tertentu sehingga dapat dikatakan nilai IG merepresentasikan tingkat penting atau tidaknya atribut untuk menentukan sebuah kategori.

Setiap atribut dihitung nilai IGnya kemudian dirangking dari nilai yang terbesar hingga yang terkecil, lalu dipilih atribut dengan nilai IG yang bukan 0. Atribut yang awalnya berjumlah 8932 atribut menjadi 394 atribut, yaitu mengalami reduksi sekitar 96% dibandingkan jumlah atribut awal namun dengan tetap menjaga hasil akurasi agar tetap baik.

B. METODE KLASIFIKASI

Algoritma C4.5 merupakan salah satu algoritma yang dapat digunakan untuk mengkonstruksi sebuah pohon keputusan yang merupakan pengembangan dari algoritma ID3 (Quinlan, 1993). Algoritma C4.5 mempunyai input *data training* dan *data testing*. *Data training* merupakan contoh data yang digunakan untuk membangun pohon keputusan yang telah diuji kebenarannya. Sedangkan *data*

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 6 Nomor. 1, April 2021
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

testing merupakan field-field data yang akan digunakan sebagai parameter dalam melakukan klasifikasi data.

Menurut Lakshmi et al. (2013), tahapan dari proses algoritme C4.5 diuraikan sebagai berikut.

1) Mempersiapkan data training.
2) Menghitung nilai entropy. Entropy merupakan ukuran ketidakpastian, yakni perbedaan keputusan terhadap nilai atribut tertentu. Semakin tinggi nilai entropy, semakin tinggi perbedaan keputusan (ketidakpastian). Nilai Entopy dihitung dengan rumus yang ditulis sebagai

$$Entropy = -\sum pi \times log2pi \; k \; i{=}1 \quad (1)$$

dengan S adalah himpunan kasus, pi adalah probabilitas yang diperoleh dari sum (ya) dibagi dengan total kasus.

3) Menghitung nilai gain. Gain merupakan salah satu langkah pemilihan atribut yang digunakan untuk memilih tes atribut setiap simpul pada pohon keputusan atau dengan kata lain gain merupakan tingkat pengaruh suatu atribut terhadap keputusan atau ukuran efektifitas suatu variabel dalam mengklasifikasikan data. Gain dihitung dengan rumus yang ditulis sebagai

$$Gain(S, A) = Entopy\ (S) - \sum |Si\ |\ |S| \times Entropy(Si\ )\ k\ i{=}1 \quad (2)$$

dengan S adalah himpunan kasus, A adalah atribut, $|Si\ |$ adalah jumah kasus pada partisi ke i, dan $|S|$ adalah jumlah kasus dalam S. Pada algoritma C4.5, nilai gain digunakan untuk menentukan variabel mana yang menjadi node dari suatu pohon keputusan. Suatu variabel yang memiliki gain tertinggi akan dijadikan node di pohon keputusan.

4) Menghitung nilai split info dengan rumus

$$SplitInfo(S,A) = -\sum Sj\ S \times log2\ Sj\ S\ k\ j{=}1 \quad (3)$$

dengan S adalah ruang sample, A adalah atribut, dan Sj adalah jumlah sample untuk atribut ke j.

5) Menentukan nilai gain ratio dengan rumus yang ditulis sebagai

$$GainRatio(S,A) = Gain(S,A)\ Split(S,A) \quad (4)$$

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 6 Nomor. 1, April 2021
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

dengan Gain(S,A) adalah information gain pada atribut (S,A), A adalah atribut, dan Split(S,A) adalah split information pada atribut (S,A).

Nilai gain ratio tertinggi akan digunakan sebagai atribut akar. Dengan demikian akan terbentuk pohon keputusan sebagai node 1.

6) Mengulangi proses ke-2 hingga semua cabang memiliki kelas yang sama. Proses percabangan akan berhenti apabila
   - semua kasus dalam simpul n mendapat kelas yang sama;
   - tidak ada variabel independen di dalam kasus yang dipartisi lagi;
   - tidak ada kasus di dalam cabang yang kosong.

Algoritma C4.5 memiliki kompleksitas waktu $O\ (m.\ n^2\ )$, dengan m ukuran data pelatihan dan *n* adalah banyak atribut.

Empat hal yang membedakan algoritma C4.5 dengan ID3 antara lain: tahan (robust) terhadap *data noise*, mampu menangani variabel dengan tipe diskrit maupun kontinu, mampu menangani variabel yang memiliki *missing value*, dan dapat memangkas cabang dari pohon keputusan.

III. HASIL DAN PEMBAHASAN

A. SKENARIO UJI COBA

1. Lakukan *preprocessing* pada *data training* dan *data testing* untuk menangani *noise, missing value, feature selection* sehingga data siap diolah.
2. Gunakan *data training* dengan *tools* WEKA 3.6.2 untuk menghasilkan model klasifikasi menggunakan metode C4.5 tanpa *feature selection*
3. Evaluasi hasil
4. Lakukan *feature selection* untuk mereduksi atribut yang ada
5. Lakukan pembentukan model lagi dan evaluasi hasil yang dihasilkan
6. Bandingkan hasil dengan dan tanpa *feature selection* menggunakan *Information Gain*.
7. Setelah model dirasa cukup baik, maka selanjutnya dilakukan klasifikasi kepada *data testing*

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 6 Nomor. 1, April 2021
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

B. PEMBAHASAN

Berdasarkan model yang dibangun menggunakan algoritma C.45 didapatkan hasil:

- Tanpa *Feature Selection*

Tabel 3 Hasil Akurasi Tanpa *Feature Selection*

Jumlah Atribut	8932
Akurasi	89.25 %

Tabel 4 Hasil *Precision dan Recall* Tanpa *Feature Selection*

Kategori	Precision	Recall
A	0.933	0.9
B	0.755	0.88
C	0.905	0.86
D	0.957	0.89
E	0.918	0.9
F	0.916	0.925
Weighted Avg.	0.897	0.893

- Dengan *Feature Selection*

Tabel 5 Hasil Akurasi dengan *Feature Selection*

Jumlah Atribut	394
Akurasi	88.58 %

Tabel 6 Hasil *Precision dan Recall* dengan *Feature Selection*

Kategori	Precision	Recall
A	0.936	0.875
B	0.743	0.91
C	0.888	0.83
D	0.958	0.905
E	0.93	0.865
F	0.903	0.93
Weighted Avg.	0.893	0.886

Dari hasil perbandingan klasifikasi dengan dan tanpa *feature selection* menggunakan *Information Gain* didapatkan hasil bahwa klasifikasi dengan *feature selection* jauh lebih baik deibandingkan tanpa *feature selection* melihat akurasi yang dihasilkan hanya menurun sebesar **0.67%** namun dengan atribut yang direduksi sebesar **96%** sehingga klasifikasi menggunakan C4.5 dengan *feature selection* menggunakan *Information Gain* dinilai efektif terutama dalam segi akurasi maupun kecepatan komputasi.

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 6 Nomor. 1, April 2021
P –ISSN : 2541-1179, E-ISSN : 2581-1711
Ojs :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

IV. KESIMPULAN

1. Pada tahap *preprocessing* dilakukan penanganan *noise, missing value*, pengurangan fitur, *stopword* dan *stemming* dengan tujuan untuk mengurangi jumlah atribut.
2. Penggunaan *Information Gain* sebagai *feature selection* dirasa sangat efektif untuk mereduksi atribut yang dianggap tidak penting yaitu sebesar **96%**.
3. C4.5 merupakan algoritma yang cukup sederhana dan efisien sebagai *classifier* untuk mengklasifikasikan dokumen, terlihat dengan hasil akurasi sebesar **88.58%**.

DAFTAR PUSTAKA

Panji Bimo Nugroho Setio, Dewi Retno Sari Saputro, dan Bowo Winarno. 2020. Klasifikasi dengan Pohon Keputusan Berbasis Algoritma C4.5. *PRISMA, Prosiding Seminar Nasional Matematika*. Surakarta: Universitas Sebelas Maret. ISSN 2613-9189. hal 64-71.

Bobby Suryo Prakoso, Didi Rosiyadi, dan Dedi Aridarma. 2019. Optimalisasi Klasifikasi Berita Menggunakan Feature Information Gain Untuk Algoritma Naive Bayes Terhubung Random Forest. Jurnal PILAR Nusa Mandiri Vol. 15. 2 September 2019. Jakarta: PPPM Nusa Mandiri. hal 211-218.

Lakshmi, T.M., A. Martin, R.M. Begum, and Dr.V.P. Venkatesan. (2013). An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data, I.J.Modern Education and Computer Science.

SP, Dedy Mayor. 2010. *Analisis dan Implementasi Deteksi EMail Spam Menggunakan Karakter N-Grams*. Bandung: IT Telkom