e-ISSN 2540-7902 dan p-ISSN 2541-366X

Penerapan Metode Naïve Bayes pada Klasifikasi Judul Jurnal

*1st Arisa Tien Hardianti
Universitas Muslim Indonesia
Fakultas Ilmu Komputer
Makassar, Indonesia
arisatien55@gmail.com

2nd Abdul Rachman Manga
Universitas Muslim Indonesia
Fakultas Ilmu Komputer
Makassar, Indonesia
abdulrachman.manga@umi.ac.id

3rd Herdianti Darwis
Universitas Muslim Indonesia
Fakultas Ilmu Komputer
Makassar, Indonesia
herdianti.darwis@umi.ac.id

***Abstrak*— Jurnal merupakan sebuah dokumen yang membahas mengenai sebuah penelitian dan berfokus pada 1 bidang kelimuan. Dalam jurnal keilmuan teknik informatika, ada banyak kategori yang bisa ditentukan, semakin banyak jurnal yang terbit, semakin banyak bidang keilmuan yang dapat dikelompokkan. Dalam penelitian ini, metode naïve bayes digunakan untuk mengklasifikasikan judul jurnal berdasarkan bidang keilmuan dalam dunia ilmu komputer yang terdapat pada data jurnal Seminar Nasional Ilmu Komputer (SNRIK) tahun 2016 dengan menghitung akurasi data. Selain itu, digunakan stopword removal untuk menghilangkan kata yang tidak memiliki arti, stemming untuk mendapatkan kata dasar dari judul jurnal yang akan dikelompokkan serta inverted index yaitu matriks antara term dan data. Dari penelitian yang dilakukan menghasilkan akurasi data sebesar 50% sehingga judul jurnal mendapatkan kategori yang sesuai.**

Kata kunci—naïve bayes; jurnal; klasifikasi;

I. PENDAHULUAN

Jurnal ilmiah merupakan salah satu bukti karya tertulis yang dihasilkan oleh seorang peneliti yang melakukan penelitian pada suatu topik atau bidang tertentu. Jurnal-Jurnal tersebut umumnya diterbitkan oleh lembaga-lembaga penyedia informasi ilmiah maupun institusi perguruan tinggi[1]. Pada dasarnya, jurnal yang telah terbit digunakan sebagai referensi guna membantu mahasiswa serta dosen dalam menyelesaikan tugas akhir maupun penelitian mereka. Pada jurnal penelitian khususnya jurnal teknik informatika, memiliki banyak bidang keilmuan, semakin banyak jurnal yang terbit, semakin banyak bidang keilmuan yang dapat dikelompokkan, diantaranya yaitu: *Artificial Intelligence* (AI), Sistem Informasi (Sisfo), Jaringan komputer, Data Mining, Sistem Pendukung Keputusan (SPK), Microcontroller, *E-Government*, Multimedia, *Machine Learning* dan lain sebagainya.

Informasi penting dari jurnal berupa kategori yang menggambarkan pokok pembahasan secara umum. Pemberian label kategori dapat membantu konsumen dalam memahami isi jurnal, tanpa harus membaca secara keseluruhan.[2]

Ketika ingin menentukan bidang keilmuan apa yang dimiliki jurnal tersebut, terkadang jurnal itu bisa masuk dari satu, dua atau bahkan lebih bidang keilmuan hanya dilihat dari judulnya saja, sehingga perlu dilakukan pengelompokan atau klasifikasi untuk menentukan kategori atau bidang keilmuan yang sesuai dengan jurnal tersebut.

Umumnya, terdapat beberapa metode klasifikasi yang bisa digunakan, seperti *Decision Tree*, *Naïve Bayes Classifier*, *Artificial Neural Network* (ANN), *k-Nearest Neighbour* dan lain sebagainya. Pada penelitian ini, metode yang digunakan yaitu *Naïve Bayes Classifier*, dimana metode tersebut akan mengklasifikasikan data jurnal sesuai bidang keilmuan tersebut.

II. METODOLOGI

A. *Pre-Processing*

Pre-processing pada proses klasifikasi bertujuan untuk membangun kumpulan term yang berisi topik yang dikandung oleh dokumen. Pembuatan inverted index mengaitkan konsep linguistic processing yang bertujuan untuk mengekstrak term-term penting.

Salah satu metode yang digunakan untuk klasifikasi yaitu metode *naïve bayes*, tapi sebelum melakukan klasifikasi, perlu dilakukan pre-processing, adapun tahap pre-processing yaitu:

- Pemisahan rangkaian term (*tokenization*). *Tokenization* adalah memisahkan kata tiap kalimat atau paragraf menjadi potongan kata tunggal atau *termword.* Pada tahan ini juga menghilangkan karakter-karakter tertentu dan mengubah huruf kapital menjadi huruf kecil.
- Stopword. Dimana kata atau term penghubung akan dihapus dari database meskipun kata tersebut sering muncul di setiap dokumen.
- Stemming. Kata-kata yang sering muncul di dalam dokumen yang memiliki arti morfologi akan diubah menjadi kata dasar.[3]

B. *Metode Naïve Bayes*

Klasifikasi merupakan proses untuk menentukan sebuah kategori dari sekumpulan objek yang kategorinya tidak diketahui. Pengkategorisasian teks menjadi suatu hal yang penting dan kebutuhannya yang semakin meningkat seiring berjalannya waktu, karena data semakin lama akan semakin bertambah.sehingga perlu digunakan metode untuk

*Corresponding Author

mengklasifikasi data uji untuk menghasilkan kategori yang sesuai[4].

Naïve Bayes Classifier (NBC) adalah metode klasifikasi yang berdasarkan probabilitas dan Teorema Bayesan dengan asumsi bahwa setiap variable X bersifat bebas (*independence*). Dengan kata lain NBC mengasumsikan bahwa keberadaan sebuah atribut (*variable*) tidak ada kaitannya dengan keberadaan atribut (*variable*) yang lain[5]. Keuntungan menggunakan metode naïve bayes adalah metode ini hanya membutuhkan jumlah data latih yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses klasifikasi. Metod naïve bayes bisa bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan[6].

Perhitungan perbandingan antara term pada data uji dengan setiap kelas yang ada dengan menggunakan persamaan (1).

$$P(a_i|v_j) = \frac{n_c + mp}{n + m} \quad (1)$$

Keterangan:

P = probabilitas

a_i = kelas atau kategori pada data latih

v_j = kata atau term yang akan diuji

n = jumlah term pada data latih dimana v = v_j

nc = jumlah term dimana v = v_j dan a = a_i

p = probabilitas setiap kelas dalam data latih

m = jumlah term pada data uji

Sedangkan untuk menentukan klasifikasi data uji digunakan persamaan (2)[7].

$$V_{nb} = argmax\ v_{j \in V} P(v_j) \prod P(a_i|v_j) \quad (2)$$

C. Confussion Matrix

Confussion matrix adalah sebuah tabel yang menyatakan jumlah data uji yang benar diklasifikasikan dan jumlah data uji yang diklasifikasikan. Contoh *confussion matrix* untuk klasifikasi biner ditunjuk pada Tabel I.

TABEL I. CONFUSSION MATRIX UNTUK KLASIFIKASI BINER

		Kelas prediksi	
		1	0
Kelas sebenarnya	1	TP	FN
	0	FP	TN

Keterangan: TP = True Possitive, yaitu jumlah dokumen dari kelas 1 yang benar dan diklasifikasikan sebagai kelas 1.

TN = True Negative, yaitu jumlah dokumen dari kelas 0 yang benar dan diklasifikasikan sebagai kelas 0.

FP = False Possitive, yaitu jumlah dokumen dari kelas 0 yang benar dan diklasifikasikan sebagai kelas 1.

FN = False Negative, yaitu jumlah dokumen dari kelas 1 yang benar dan diklasifikasikan sebagai kelas 0.

Perhitungan akurasi dinyatakan dalam persamaan (3)[8].

$$Akurasi = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% \quad (3)$$

III. HASIL DAN PEMBAHASAN

A. Pre-Processing

Pre-processing dilakukan pada data jurnal yang telah diklasifikasi dan menghasilkan kategori secara manual seperti ditunjukkan pada Tabel II.

TABEL II. KOLEKSI DATA JURNAL

No.	Data Jurnal	Kategori
D1	Big Data Analisis Sebuah Peluang dan Tantangan Lulusan Informatika dalam Menghadapi Pasar Global di Indonesia Timur	Data Mining
D2	Klasifikasi Nasabah Kredit Bank Menggunakan Algoritma K-Nearest Neighbour berbasis Forward Selection	Artificial Intelligence
D3	Adaptive Neuro Fuzzy Inference System Anfis berbasis Fitur Seleksi Relief untuk Prediksi Harga Minyak Kelapa	Artificial Intelligence
…	…	…
…	…	…
D15	Pemanfaatan NLP dan Information Base untuk E-Government Tingkat Kecamatan Kota Makassar Berbasis SMS	E-Government

Pada tabel 2 terdapat 15 data jurnal yang mempunyai masing-masing kategori yakni 1 data kategori data *mining*, 3 data kategori *Artificial Intelligence*, 2 data kategori mikrokontroler, 3 data kategori sistem informasi, 1 data kategori sistem terdistribusi, 3 data kategori jaringan, 1 kategori sistem informasi geografis dan 1 data kategori *e-government.*

Selanjutnya masuk pada tahap *casefolding*, berikut hasil casefolding ditunjukkan pada Tabel III.

Tabel III menunjukkan hasil casefolding dimana data tersebut akan menghilangkan karakter khusus dan mengubah huruf kapital menjadi huruf kecil. Selanjutnya yaitu stopwords, dimana kata penghubung dari setiap jurnal akan dihapus. Hasil dari stop words ditunjukkan pada Tabel IV.

TABEL III. HASIL CASEFOLDING

No.	Data Jurnal	Kategori
D1	big bata analisis sebuah peluang dan tantangan lulusan informatika dalam menghadapi pasar	Data Mining

No.	Data Jurnal	Kategori
	global di indonesia timur	
D2	klasifikasi nasabah kredit bank menggunakan algoritma k nearest neighbour berbasis forward selection	Artificial Intelligence
D3	adaptive neuro fuzzy inference sistem anfis berbasis fitur seleksi relief untuk prediksi harga minyak kelapa	Artificial Intelligence
…	…	…
…	…	…
D15	pemanfaatan nlp dan information base untuk e government tingkat kecamatan kota makassar berbasis sms	E-Government

TABEL IV. HASIL STOPWORDS

No	Data Jurnal
D1	big bata analisis sebuah peluang tantangan lulusan informatika menghadapi pasar global indonesia timur
D2	klasifikasi nasabah kredit bank menggunakan algoritma k nearest neighbour berbasis forward selection
D3	adaptive neuro fuzzy inference sistem anfis berbasis fitur seleksi relief prediksi harga minyak kelapa
…	…
…	…
D15	pemanfaatan nlp information base e government tingkat kecamatan kota makassar berbasis sms

Setelah melalui tahap stopwords, tahap terakhir yaitu stemming, dimana data jurnal tersebut akan diubah menjadi kata dasar, hasil dari stemming data jurnal ditunjukkan pada Tabel V.

TABEL V. HASIL STEMMING

No.	Data Jurnal
D1	big data analisis luang tantang lulus informatika hadap pasar global indonesia timur
D2	klasifikasi nasabah kredit bank guna algoritma k nearest neighbor basis forward selection
D3	adaptive neuro fuzzy inference sistem anfis basis fitur seleksi relief prediksi harga minyak kelapa
…	…
…	…
D15	manfaat nlp information base e government tingkat kecamatan kota makassar basis sms

Pada tahap pre-processing diperoleh inverted index yaitu matriks antara term dan data, berikut hasil inverted index dari data jurnal pada Tabel VI.

Tabel VI merupakan matriks hubungan antara term dan data. Tahap selanjutnya yaitu melakukan proses klasifikasi dengan metode *naïve bayes*.

TABEL VI. INVERTED INDEX DARI DATA UJI

Term	D1	D2	D3	…	…	D15
Big	1	0	0	…	…	0
Data	1	0	0	…	…	0
analisis	1	0	0	…	…	0
…	…	…	…	…	…	…
Sms	0	0	0	…	…	1

B. *Naïve Bayes Classifier*

Tahap pertama dengan menghitung probabilitas masing-masing kategori terhadap data latih sebanyak 15 data pada tabel 2. Diketahui bahwa dari 15 data latih ada 8 kategori, yaitu data *mining*, *artificial intelligence*, sistem informasi, mikrokontroler, sistem informasi geografis, jaringan komputer, sistem terdistribusi, dan *e-government.*

Probabilitas disimbolkan sebagai p, persamaan (4).

$$p(data\ mining) = \frac{jumlah\ kelas\ data\ mining}{jumlah\ data\ latih} \quad (4)$$

$$\pi(\delta\alpha\tau\alpha\ \mu\iota\nu\iota\nu\gamma) = 1/15 = 0.067$$

Berdasarkan perhitungan probabilitas untuk kategori data mining diperoleh nilai probabilitas yaitu 0.067, *artificial intelligence* 0.2, sistem informasi 0.2, mikrokontroler 0.13, sistem terdistribusi 0.067, jaringan 0.2, sistem informasi geografis 0.067 dan *e-government* 0.067.

Selanjutnya dilakukan klasifikasi data uji menggunakan metode NBC, berikut data jurnal yang akan diuji pada Tabel VI.

TABEL VII. DATA JURNAL

No.	Data Jurnal	kategori
D16	Sistem Transportasi Cerdas Sebagai Solusi Kota Tumbuh Pesat di Negara Berkembang	?
D17	Ujian Online Mahasiswa Ilmu Komputer Berbasis Smartphone	?
D18	Protokol Arsitektur Software Defined Networking SDN	?
D19	Implementasi provisioning Indihome Berbasis FTTH Fiber to the Home dengan teknologi GPON Gigabit Passive Optical Network di PT Telkom Akses	?

Pada Tabel VII terdapat 4 data uji yang belum memiliki kategori, data uji akan melalui tahapan pre-processing. Berikut hasil pre-processing ditunjukkan pada Tabel VIII.

Selanjutnya masuk ketahap pengujian data jurnal, langkah awal yaitu menentukan nilai n, nc, p, dan m, berikut nilai-nilai untuk kategori “data mining” ditunjukkan pada Tabel IX.

TABEL VIII. HASIL PRE-PROCESSING PADA DATA UJI

No.	Data Jurnal	kategori
D16	sistem transportasi cerdas solusi kota tumbuh pesat negara kembang	?
D17	ujian online mahasiswa ilmu komputer berbasis smartphone	?
D18	protokol arsitektur software defined networking sdn	?
D19	implementasi provisioning indihome basis ftth fiber to the home teknologi gpon gigabit passive optical network pt telkom akses	?

TABEL IX. NILAI DATA UJI

term	n	nc	p	M
"sistem"	15	0	0.067	9
"transportasi"	15	0	0.067	9
"cerdas"	15	0	0.067	9
"solusi"	15	0	0.067	9
"kota"	15	0	0.067	9
"tumbuh"	15	0	0.067	9
"pesat"	15	0	0.067	9
"negara"	15	0	0.067	9
"kembang"	15	0	0.067	9

Dari nilai-nilai tersebut dengan menggunakan persamaan 1 diperoleh perhitungan sebagai berikut:

$$P(data\ mining\ |sistem)\ \frac{0+(9.0{,}067)}{15+9}=0{,}025125 \quad (1)$$

Untuk perhitungan $P(a_i|v_j)$ yang lainnya dilakukan proses yang sama untuk setiap term disetiap kategori. Selanjutnya, dengan menggunakan persamaan 2 yaitu mencari nilai maksimal dari hasil perkalian dari nilai probabilitas.

V(data mining) = 0.067*0.025*0.025*0.025*0.025* 0.025*0.025*0.025*0.025*0.025 = $2{,}6731\text{x}10^{-16}$

V(artificial intelligence) = 0.2*0.075*0.075*0.075*0.075* 0.075*0.075*0.075*0.075*0.075= $1{,}5016\text{x}10^{-11}$

V(sistem terdistribusi) = 0.067*0.0667*0.025*0.025* 0.025*0.025*0.025*0.025*0.025*0.025= $7{,}1706\text{x}10^{-16}$

V(sistem informasi) = 0.2*0.0158*0.075*0.075*0.075* 0.075*0.075*0.075*0.075*0.1166= $4{,}93\text{x}10^{-11}$

V(jaringan komputer) = 0.2*0.01166*0.075*0.075*0.075* 0.075*0.075*0.075*0.075*0.075= $2{,}335\text{x}10^{-11}$

V(Mikrokontroler) = 0.13*0.0904*0.04875*0.04875* 0.04875*0.04875*0.04875*0.04875*0.04875*0.04875* 0.04875= $3{,}749\text{x}10^{-13}$

V(Sistem informasi geografis) = 0.067*0.0667*0.025* 0.025*0.025*0.0667*0.025*0.025*0.025*0.025 =$1{,}889\text{x}10^{-15}$

V(E-government) = 0.067*0.025*0.025*0.025*0.025* 0.0667*0.025*0.025*0.025*0.025 = $7{,}1063\text{x}10^{-16}$

Vnb = argmax($2{,}673\text{x}10^{-16}$ | $1{,}5016\text{x}10^{-11}$ | $7{,}1706\text{x}10^{-16}$ | $4{,}93\text{x}10^{-11}$ | $2{,}335\text{x}10^{-11}$ | $3{,}749\text{x}10^{-13}$ | $1{,}889\text{x}10^{-15}$ | $7{,}1063\text{x}10^{-16}$)

Vnb = $4{,}93\text{x}10^{-11}$

Sehingga didapatkan hasil bahwa jurnal D16 memiliki kategori sistem informasi.

C. *Confussion Matrix*

Langkah terakhir yaitu menghitung akurasi data yang telah diuji dengan menggunakan confussion matrix seperti pada Tabel X.

TABEL X. CONFUSSION MATRIX DARI DATA UJI

Nomor jurnal	Kategori tanpa NBC	Kategori menggunakan NBC
D16	Artificial intelligence	Sistem informasi
D17	Sistem Informasi	Artificial intelligence, Sistem informasi
D18	Jaringan komputer	Jaringan komputer, sistem informasi, artificial intelligence
D19	Jaringan computer	Artificial intelligence

$$Akurasi=\frac{2+0}{2+0+2+0}\times 100\%=\frac{2}{4}\times 100\%=50\% \quad (2)$$

Dari perhitungan di atas disimpulkan bahwa hasil akurasi dari pengujian 4 data di atas menghasilkan nilai akurasi sebesar 50%.

IV. KESIMPULAN

Dari hasil pengujian 4 data judul jurnal menggunakan metode *naïve bayes* serta menggunakan *confussion matrix,* tingkat akurasi yang dihasilkan sebesar 50%. Pada penggunaan metode ini, nilai akurasi yang dihasilkan berpengaruh pada banyaknya data yang diuji. Untuk penelitian selanjutnya, data uji tersebut bisa menghasilkan nilai presisi, *recall* dan *error rate* serta dapat membandingkan hasil pengujian data menggunakan metode naïve bayes dengan metode klasifikasi lainnya.

DAFTAR PUSTAKA

[1] A. S. Sri Widaningsi, "Klasifikasi Jurnal Ilmu Komputer Berdasarkan Pembagian," *Semin. Nas. Teknol. Inf. dan Komun. 2018 (SENTIKA 2018)*, vol. 2018, no. Sentika, pp. 23–24, 2018.

[2] A. Indranandita, B. Susanto, and A. Rahmat, "Sistem Klasifikasi Dan Pencarian Jurnal Dengan Menggunakan Metode Naive Bayes Dan Vector Space Model," *J. Inform.*, vol. 4, no. 2, 2011.

[3] U. M. S. Notebox, "Information Retrieval & Klasifikasi Teks," pp. 1–27.

[4] A. Setiawan, I. Fitri Astuti, and A. Harsa Kridalaksana, "Klasifikasi Dan Pencarian Buku Referensi Akademik Menggunakan Metode Naïve Bayes Classifier (Nbc) (Studi Kasus: Perpustakaan Daerah Provinsi Kalimantan Timur)," *J. Inform. Mulawarman*, vol. 10, no. 1, 2015.

[5] T. F. Abidin, S. Si, and M. Tech, "Na ï ve Bayesian Classifier," 2014.

[6] S. A. Pattekari and A. Parveen, "Prediction system for heart disease using Naive Bayes," *Int. J. Adv. Comput. Math. Sci. ISSN*, vol. 3, no. 3, pp. 2230–9624, 2012.

[7] E. Meisner, "Naive Bayes Classifier example Car theft Example," *Comput. Linguist.*, vol. 36, no. 2, pp. 2–3, 2003.

[8] A. Indriani, "Klasifikasi Data Forum dengan menggunakan Metode Naïve Bayes Classifier," *Semin. Nas. Apl. Teknol. Inf.*, pp. 5–10, 2014.