Journal of Computer and Information System (J-CIS)
Vol.5, No.2, Oktober 2022, Page 1-12
ISSN (print): 2622-5859, ISSN (online): 2622-0881
DOI: https://doi.org/10.31605/jcis.v5i2

Pengelompokan Judul Penelitian Mahasiswa menggunakan Algoritma *Naïve Bayes* pada Program Studi Teknik Informatika

Kartika Sari*[1], Nurdina Rasjid[2], Adi Heri[3]
[1,2,3]Program Studi Teknik Informatika, Universitas Sulawesi Barat
E-mail: *[1]kartikasariusb@gmail.com, [2]nurdinarasyid@unsulbar.ac.id, [3]adiheri@unsulbar.ac.id

Abstrak

Judul tugas akhir bagi mahasiswa di program studi Teknik Informatika di Universitas Sulawesi Barat (UNSULBAR) terdiri dari tiga konsentrasi utama, yaitu Sistem Cerdas (Smart System), Rekayasa Perangkat Lunak (Software Engineering), dan Jaringan Komputer (Internet of Things). Namun, pengelompokan judul skripsi oleh mahasiswa di UNSULBAR belum terklasifikasi secara optimal sesuai dengan bidang konsentrasinya. Oleh karena itu, diperlukan metode pengelompokan yang dapat membantu mahasiswa dalam menemukan judul skripsi yang relevan dengan konsentrasi studinya. Salah satu teknik klasifikasi yang dapat digunakan adalah text mining, yaitu teknik data mining yang mencari pola menarik dari kumpulan data teks yang besar. Dalam penelitian ini, algoritma naive bayes digunakan untuk mengklasifikasikan judul skripsi berdasarkan topiknya. Hasil pengujian menunjukkan bahwa tingkat akurasi algoritma naive bayes dalam mengklasifikasikan judul skripsi mencapai 97% dengan nilai presisi 0,97%, recall 0,96%, dan laju error 0,03%. Dari hasil pengujian ini, dapat disimpulkan bahwa implementasi metode naive bayes dapat memilah judul skripsi ke dalam kategori kelas secara efektif dan efisien.

Kata kunci*—Naïve bayes, Data Mining, Klasifikasi Skripsi*

Abstract

The title of the thesis made by students, especially in the Informatics program at Unsulbar, consists of three concentrations: Smart Systems, Software Engineering, and Computer Networks (Internet of Things). However, the classification of thesis titles at Unsulbar, especially in the Informatics program, has not been maximally classified according to the concentration and development of informatics science that continues to evolve with the times. This is expected to help provide an overview for students to find titles that are relevant to their concentration. To classify or group the topics of student theses in the Informatics department, the thesis titles made by the students were observed. There are many classification techniques known, and one of them is text mining, which is a data mining variation that attempts to determine interesting patterns from a large amount of textual data. The results of implementing the Naive bayes algorithm were successful in classifying thesis titles using Naive bayes algorithm, which can be used to group thesis titles according to their topic class. The accuracy level obtained from the testing of 150

training data and 100 testing data was 97%. The precision test obtained a value of 0.97%, the recall got a value of 0.96%, and the error rate value was 0.03%. Based on the test results, it can be seen that the implementation of the Naive bayes method in classifying thesis titles into category classes is good, as can be seen from the accuracy level obtained from the 100 testing data and 150 training data.

Keywords—*Naïve bayes, Data Mining, Thesis classification*

1. PENDAHULUAN

Di era teknologi informasi yang kita hadapi saat ini, peran teknologi informasi menjadi sangat krusial dan diperlukan di berbagai sektor, termasuk dalam dunia akademik. Dalam perkembangan teknologi ini, jumlah dokumen elektronik yang tersimpan di perpustakaan universitas telah menjadi sangat besar. Versi digital dari berbagai karya ilmiah yang dihasilkan oleh mahasiswa dan staf akademik, seperti skripsi, laporan penelitian, laporan kerja praktik, dan lain sebagainya, sekarang dapat diakses dengan mudah. Namun, dokumen elektronik seringkali tidak mengandung banyak informasi atau pengetahuan yang dapat diekstraksi. Hal ini menyebabkan penumpukan dokumen yang membutuhkan sumber daya atau ruang penyimpanan yang signifikan [1].

Salah satu syarat untuk menyelesaikan pendidikan di suatu perguruan tinggi adalah memiliki skripsi. Seperti halnya program studi informatika di Fakultas Teknik Universitas Sulawesi Barat (UNSULBAR), semua mahasiswa diwajibkan untuk menyelesaikan penelitian yang dikemas dalam skripsi. Untuk mendapatkan gelar sarjana, mahasiswa harus menyelesaikan skripsi yang menunjukkan kemampuan mereka untuk menyusun dan menulis karya ilmiah yang relevan dengan bidang studi mereka. Skripsi adalah salah satu persyaratan akademik di universitas. Tugas akhir, menurut Syahdrajat (2015), adalah dokumentasi penting di perguruan tinggi yang dapat digunakan sebagai sumber informasi dan pembelajaran bagi setiap anggota civitas akademik [2]. Tugas akhir bermanfaat karena menggabungkan berbagai pengetahuan dengan sejawat dan rekan-rekan untuk memberikan sumbangan pada pendidikan, yang bermanfaat bagi pembaca, adik-adik kelas di perguruan tinggi, dan generasi mahasiswa di masa depan. Tugas akhir, menurut Barnawi (2015), adalah karya tulis ilmiah hasil penelitian pustaka atau lapangan yang harus dipresentasikan di depan penguji sebagai salah satu syarat untuk mendapatkan gelar sarjana (Strata-1). [3].

Di program studi Teknik Informatika Universitas Sulawesi Barat, terdapat tiga konsentrasi jurusan: Sistem Cerdas (Smart System), Rekayasa Perangkat Lunak (Software Engineering), dan Jaringan Komputer (Internet of Things). Diharapkan penelitian ini dapat membantu siswa menemukan judul yang sesuai dengan bidang konsentrasi mereka. Penelitian ini menggunakan algoritma *Naive bayes* untuk mengklasifikasikan judul skripsi berdasarkan konsentrasi di Jurusan Teknik Informatika UNSULBAR.

Data mining atau penambangan adalah proses yang diperlukan untuk pengolahan data yang besar. Data mining adalah proses mengidentifikasi pola, tren, dan wawasan dari kumpulan data yang sangat besar dengan menggunakan teknik komputasi dan statistik. Ini memerlukan ekstraksi informasi dari sekumpulan data yang sangat besar untuk menemukan pola dan hubungan yang sebelumnya tidak diketahui. Teknik penambangan data dapat digunakan untuk membuat prediksi dan menghasilkan wawasan yang dapat ditindaklanjuti dari data dari berbagai domain, seperti bisnis, keuangan, kesehatan, dan sains. Beberapa teknik penambangan data yang umum termasuk pengelompokan, klasifikasi, penambangan aturan asosiasi, dan deteksi anomali. Data mining digunakan untuk mendapatkan informasi relevan dan berguna dari kumpulan data yang kompleks dan besar [4]. *Teks mining* merupakan salah satu teknik yang digunakan untuk melakukan klasifikasi dokumen dimana teks mining merupakan variasi data mining yang

berusaha menetukan pola menarik dari sekumpulan data tekstual yang berjumlah besar [5]. *Text mining* juga disebut sebagai teknik dalam pengolahan data yang menggunakan algoritma dan teknik analisis statistik untuk mengekstrak informasi dan pengetahuan dari dokumen atau teks yang besar dan kompleks [6]. Metode klasifikasi membagi data ke dalam kategori atau kelas tertentu berdasarkan karakteristiknya. Metode ini banyak digunakan dalam berbagai bidang, seperti pengenalan citra, analisis teks, dan pengambilan keputusan [7]. *Naive bayes* adalah salah satu algoritma yang paling sering digunakan dalam teknik klasifikasi karena memanfaatkan probabilitas munculnya fitur untuk melakukan klasifikasi data. Dalam aplikasinya, teknik klasifikasi dapat membantu pengambilan keputusan, pengolahan data, dan mengoptimalkan kinerja sistem [8].

Naive bayes adalah salah satu algoritma klasifikasi yang paling umum digunakan. Ini adalah salah satu teknik data mining yang menggunakan teori probabilitas untuk menentukan kelas data. Metode ini didasarkan pada teorema Bayes, suatu rumus matematika yang menghitung kemungkinan suatu peristiwa terjadi berdasarkan kemungkinan peristiwa yang terkait dengannya [9].

Hasil penelitian ini diharapkan dapat membantu mahasiswa dalam menentukan judul skripsi yang sesuai dengan konsentrasi mereka. Selain itu, penelitian ini dapat berkontribusi pada pengembangan sistem informasi akademik, khususnya berkaitan dengan bagaimana topik skripsi mahasiswa dikelompokkan.

2. METODE

Penelitian ini menggunakan metode kuantitatif, yang berarti bahwa peneliti mencari pengetahuan dengan memberikan data dalam bentuk angka dan kemudian menggunakan angka tersebut untuk melakukan analisis keterangan. Secara sederhana, penelitian kuantitatif adalah penelitian ilmiah yang disusun secara sistematis terhadap bagian-bagian dan dilakukan untuk menemukan kausalitas antara bagian-bagian tersebut. bahwa penelitian dilakukan untuk menggunakan teori tertentu untuk melakukan uji coba terhadap masalah tertentu sehingga hasil uji coba yang tepat antara masalah yang diuji dan teori yang digunakan.

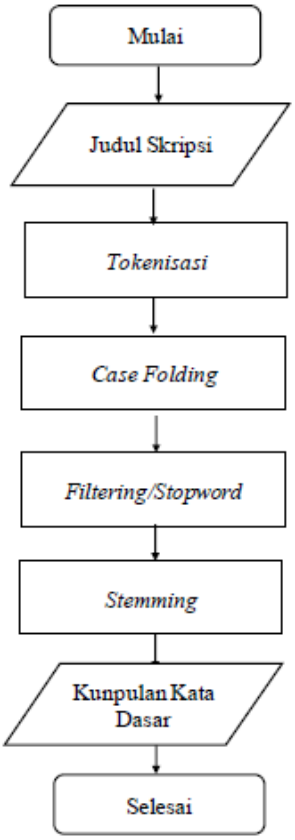
2.1 Pengambilan Data

Data dikumpulkan dari Fakultas, terutama dari program studi Informatika Unsulbar. File judul skripsi adalah data penting yang dimaksud, yang akan digunakan sebagai data uji coba untuk mengevaluasi keberhasilan sistem yang saat ini digunakan untuk mengorganisasi dokumen skripsi. Data skripsi yang digunakan dalam penelitian ini berasal dari tahun akademik 2009–2018.

2.2 Pre-processing

Preprocessing memiliki tujuan untuk mempersiapkan data agar dapat dengan mudah dianalisis. Proses *preprocessing* terdiri dari beberapa langkah, termasuk *case folding, tokenizing*, dan *filtering*. *Case folding* dilakukan untuk mengubah semua karakter huruf dalam teks menjadi huruf kecil, dengan hanya menerima karakter 'a' hingga 'z'. *Tokenizing* merupakan proses pemotongan string input menjadi kata-kata yang membentuknya. *Filtering* dilakukan dengan mengambil kata-kata yang penting dari hasil *tokenizing* dan menghapus *stopwords*. *Stopwords* merupakan kata-kata yang tidak memberikan deskripsi yang signifikan, sehingga dapat dihilangkan tanpa mempengaruhi proses analisis. Contoh *stopwords* dalam bahasa Indonesia adalah "yang", "dan", "dari", "di", "seperti", dan lain sebagainya.

Tahap *preprocessing* ini merupakan langkah awal dalam pengolahan data. Pada kasus ini, data yang digunakan adalah file judul skripsi, sehingga tahap *preprocessing* sangat penting untuk proses pengolahan data selanjutnya. Seperti yang telah dijelaskan sebelumnya, tahap *preprocessing* terdiri dari beberapa subproses, termasuk *case folding*, tokenisasi, *filtering/stopword*, dan *stemming*.



Gambar 1 Alur Proses *Preprocessing*

a) Tokenisasi

Proses tokenisasi menghilangkan tanda baca, angka, dan semua karakter selain alphabet. Setelah itu, semua karakter selain huruf alfabet dihilangkan. Langkah berikutnya adalah mengubah semua huruf kapital menjadi huruf kecil, proses yang dikenal sebagai case folding. Selanjutnya, mengubah setiap kata dari dokumen sebelumnya menjadi satu kata, atau token.

b) Case Folding

Langkah selanjutnya melakukan perubahan semua huruf kapital menjadi huruf kecil atau yang biasa disebut *case folding.*

c) Filtering/Stopword

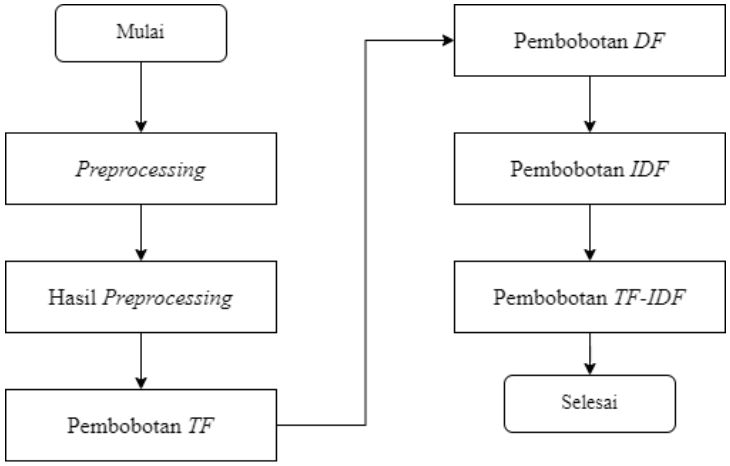
Proses selanjutnya, *filtering/stopword*, menghilangkan atau menyaring kata-kata yang dapat menggambarkan isi dokumen. Untuk menentukan kata atau *term* mana yang harus dihilangkan dan tidak berdampak pada proses, gunakan kamus *stopword* yang sudah ada. Contoh frasa penutup dalam bahasa Indonesia adalah "yang", "dan", "dari", "di", "seperti", dan sebagainya.

d) Stemming

Proses *stemming* menghasilkan kata dasar atau kata inti dari hasil *filtering*. Pada tahap ini, kata imbuhan awalan dan akhiran dihilangkan, yang menghasilkan kata dasar dari setiap kata, atau token.

2.3 Pembobotan Kata

Pembobotan adalah proses mengubah kata menjadi bentuk angka atau vektor. *TF* adalah frekuensi dari kemunculan *term* dalam dokumen yang bersangkutan. Semakin besar jumlah kemunculan suatu *term* dalam dokumen, maka akan semakin besar bobotnya atau akan memberikan nilai kesesuaian yang semakin besar. Sedangkan *TF-IDF* merupakan metode untuk menghitung nilai *TF* dan *IDF* pada setiap *token* (kata) disetiap dokumen. Berikut proses pembobotan *TF-IDF* yang terdapat pada gambar 2 dibawah ini.



Gambar 2 Diagram Alur Pembobotan *TF-IDF*

Gambar 2 adalah setelah proses hasil *preprocessing* didapatkan selanjutnya akan diproses ketahap pembobotan kata atau *term*. Pembobotan ini dilakukan untuk mengetahui nilai dari setiap *term* yang mewakili isi dokumen. Untuk alur proses dari pembobotan kata ditunjukkan pada gambar 2 diatas.

2.4 Naïve Bayes

Metode *Naive bayes* telah terbukti efektif dalam beberapa aplikasi, seperti klasifikasi email spam, analisis sentimen, dan pengenalan teks. Ini karena metode ini disebut "naive" karena menganggap bahwa semua atribut dalam data independen satu sama lain, sehingga tidak memperhitungkan hubungan antar atribut yang mungkin ada. Dalam *Teorema Bayes*, suatu probabilitas bersyarat dinyatakan sebagai persamaan berikut [10].

$$P(H|X) = \frac{P(H|X)P(H)}{P(X)} \qquad (1)$$

Berikut penjelasan dari persamaan diatas:
X = Data dengan *class* yang belum diketahui
H = Hipotesis pada data X yang merupakan suatu *class* khusus
$P(H|X)$ = Nilai probabilitas pada hipotesis H berdasarkan kondisi X
$P(H)$ = Nilai probabilitas pada hipotesis H
$P(X|H)$ = Nilai probabilitas pada hipotesis X berdasarkan kondisi H
$P(X)$ = Nilai probabilitas pada X

Dimana X adalah bukti, H adalah hipotesis, $P(H|X)$ adalah probabilitas bahwa hipotesis H benar untuk bukti X atau dengan kata lain $P(H|X)$ merupakan probabilitas *posterior H* dengan syarat X. $P(X|H)$ adalah probabilitas bahwabukti X benar untuk hipotesis H atau probabilitas

posterior X dengan syarat *H*, *P(H)* adalah probabilitas *prior* hipotesis *H*, dan P(X) adalah probabilitas *prior* bukti *X*.

Dalam bidang *machine learning*, *X* adalah sebuah *tuple* atau objek data, *H* adalah hipotesis atau dugaan bahwa *tuple X* adalah kelas *C*. Secara spesifik dalam masalah klasifikasi *P (H | X)* sebagai probabilitas bahwa hipotesis H benar untuk *tuple* X atau dengan kata lain *P (H | X)*adalah probabilitas bahwa *tuple X* berada dalam kelas *C*. Sementara itu, *P(H)* adalah probabilitas *prior* bahwa hipotesis *H* benar untuk setiap *tuple* tidak peduli nilai-nilai atributnya sedangkan *P(X)* adalah probabilitas *prior* dari *tuple X*.

Proses penggunaan model *Naive bayes classifier* himpunan data *training set* dimisalkan sebagai D, yang berisi sejumlah *tuple* X. Kelasnya dimisalkan sebagai *m*, yaitu ($C1$, $C2$, … Cm) adalah data bertipe kategorial untuk sebuah *tuple* masukan X, setiap *tuple* adalah berdimensi *n* yang dinyatakan sebagai $A1$, $A2$, … An. Probabilitas *prior* masing – masing kelas dihitung sehingga dihasilkan sebuah tabel kelas, dimana P(Ci) = jumlah *tuple* dikelas Ci dibagi total semua *tuple* dalam himpunan data latih D.

Hitung probabilitas setiap nilai pada semua *n* atribut ($A1$, $A2$, … An) untuk seluruh *m* kelas ($C1$, $C2$, … Cm) sehingga dihasilkan sejumlah *n* tabel probabilitas, jika ada probabilitas bernilai 0, maka dilakukan *Laplacian correction*. *Laplacian correction* atau *additive smoothing* adalah suatu cara untuk menangani nilai probabilitas 0 (nol). Dari sekian banyak data di *training set*, pada setiapperhitungan datanya ditambah 1 (satu) dan tidak akan membuat perbedaan yang berarti pada estimasi probabilitas sehingga bisa menghindari kasus nilai probabilitas 0 (nol). Sehingga dihasilkan sebanyak (*n* + 1) tabel, yaitu : satu tabel probabilitas kelas dan *n* tabel berisi probabilitas setiap nilai pada semua atribut yang ada. Persamaan *Laplacian correction* dapat dilihat pada persamaan berikut.

$$P(Xkj \mid Ci = \frac{f(Xkj \mid Ci+1}{(Ci)+ \mid W \mid} \qquad (2)$$

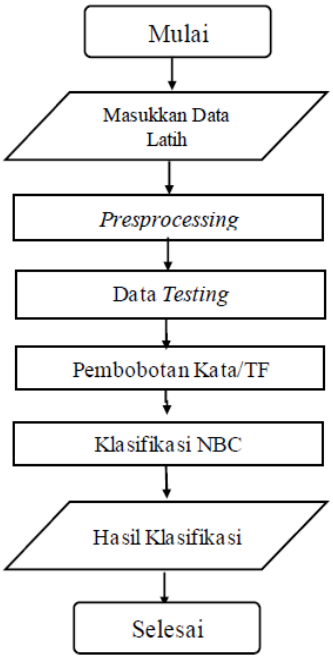
Dimana;

$P(Xkj \mid Ci)$ = peluang kata Xkj pada kategori Ci,
$f(Xkj \mid Ci)$ = nilai frekuensi kemunculan kata Xkj pada kategori Ci,
$f(Ci)$ = jumlah frekuensi kemunculan kata pada kategori Ci,
$|W|$ = jumlah keseluruhan atribut kata (*token*) yang digunakan.

Untuk memprediksi label kelas dari *tuple* X, maka harus dihitung probabilitas *P(X | Ci) P(Ci)* untuk setiap kelas Ci. Selanjutnya mencari kelas Ci yang menghasilkan probabilitas P(X | Ci) *P(Ci)* maksimum sebagai kelas keputusan. Persamaan berikut dapat digunakan untuk mengklasifikasikan sebuah *tuple* masukan dengan mencari probabilitas maksimum dari semua kelas yang ada [11].

2.5 Menentukan atribut kata

Atribut kata, juga dikenal sebagai token, ditentukan berdasarkan hasil hitung frekuensi atau kemunculan kata yang benar-benar menunjukkan isi judul skripsi. Ada kemungkinan bahwa kata tersebut menunjukkan kelas tertentu karena seringnya muncul. 10 kata yang paling sering digunakan dan relevan dipilih sebagai karakteristik kata. Setelah proses *preprocessing*, proses selanjuntya menghitung pembobotan kata yang sudah dijelaskan di atas. Setelah proses pembobotan kata dan hasilnya diketahui, proses perhitungan bobot kemungkinan setiap *term* yang termasuk dalam judul skripsi dimulai. menggunakan *naive bayes* untuk menghasilkan hasil klasifikasi judul berdasarkan kategori kelasnya. Alur sistem secara umum dapat dilihat pada gambar 3.



Gambar 3 Alur Sistem Secara Umum

2.6 Pengujian

Pengujian dilakukan dengan memanfaatkan tabel acuan yang dikenal sebagai *"confusion matrix"*. Secara dasar, *confusion matrix* memberikan informasi perbandingan antara hasil klasifikasi yang dilakukan oleh sistem (model) dengan hasil klasifikasi yang sebenarnya. *Confusion matrix* ini berbentuk tabel matriks yang menggambarkan kinerja model klasifikasi pada serangkaian data uji, di mana nilai sebenarnya dari data tersebut diketahui.

Table 1 *Confusion matrix*

Klasifikasi Data Benar	Diklasifikasi Sebagai +(0)	Diklasifikasi Sebagai -(1)
+(0)	*True Positive*	*False Negative*
-(1)	*False Positive*	*True Negative*

Dimana;
True Positive (TP) : data positif yang diprediksi benar
True Negative (TN) : data negatif yang diprediksi benar.
False Postive (FP) : data negatif namun diprediksi sebagai data positif.
*False Negative (FN***)** : data positif namun diprediksi sebagai data *negative.*

Sehingga diturunkan rumus sebagaiberikut:

1) Rumus Akurasi

Accuracy = (TP + TN) / (TP+FP+FN+TN) (3)

2) *Rumus Precission*

Rumus Precission : (TP) / (TP + FP) (4)

3) *Rumus Recall*

Recall = (TP) / (TP + FN) (5)

3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan data judul skrpsi prodi informatika fakultas teknik universitas sulawesi barat mulai dari angkatan 2009-2018 yang diambil dari sintaks informatika untuk melakukan klasifikasi atau pengelompokan judul skripsi berdasarkan kategori kelasnya yaitu RPL, Sistem Cerdas dan IoT. Data yang digunakan dalam penelitian ini sebanyak 250 judul yang dibagi kedalam data training sebanyak 150 judul sedangkan untuk data *testing* sebanyak 100 judul. Tabel dibawah ini berisikan *dataset* yang digunakan dan telah diberi kelas kategori.

Tabel 2 Contoh Data Training

No	Judul	Kategori
1	Sistem Informasi Pengolahan Data Simpan Pinjam Koperasi Pegawai Republik Indonesia Karya Dharma Kabupaten Majene	RPL
2	Implementasi *Elimination and Choice Translation Reality* (*Electre*) Untuk Menentukan Calon Pembimbing Tugas Akhir Mahasiswa	SistemCerdas
3	Sistem Otomasi Pada *Smart Home* Berbasis *Internet Of Things* (*Iot*)	*IoT*

(sumber: Data Program Studi Teknik Informatika Unsulbar)

3.1 Tokenisasi

Proses pengubahan teks menjadi kata-kata terpisah, yang dikenal sebagai tokenisasi, bertujuan untuk memisahkan atau memecah teks atau kalimat menjadi sejumlah kata individu. Hal ini dilakukan agar memudahkan dalam memberikan bobot pada setiap kata. Berikut adalah hasil tokenisasi untuk data training judul skripsi dapat dilihat pada tabel 3.

Tabel 3 Hasil *Tokenisasi* Judul skirpsi

No.	Judul	Hasil *Case Folding*	Tokenisasi	Kategori
1	Sistem Informasi Pengolahan Data Simpan Pinjam Koperasi Pegawai Republik Indonesia Karya Dharma Kabupaten Majene	Sistem informasi pengolahan data simpan pinjam koperasi pegawai republic Indonesia karya dharma kabupaten majene	Sistem Informasi Pengolahan Data Kabupaten *Dan seterusnya…*	RPL
2	Implementasi Elimination and Choice Translation	implementasi elimination and choice translation reality (electre) untuk menentukan calon pembimbing tugas akhir mahasiswa	Implementasi *Elimination And Choice Reality* *Dan seterusnya…*	SistemCerdas
3	SistemOtomasi Pada Smart Home Berbasis Internet Of Things (*Iot*)	Sistem otomasi pada smart home berbasis internet of things (*Iot*)	System Otomasi pada *smart home* Berbasis *intermet of things* *Dan seterusnya...*	*IoT*

3.2 Case Folding

Proses case folding merupakan langkah dalam menyamakan bentuk huruf dengan mengubah semua huruf menjadi huruf kecil. Tabel 4 adalah hasil data yang telah dilakukan proses *case folding:*

Tabel 4 Hasil *Case Folding* Judul skripsi

No	Judul	Hasil *Case Folding*	Kategori
1	Sistem Informasi Pengolahan Data Simpan Pinjam Koperasi Pegawai Republik Indonesia Karya Dharma Kabupaten Majene	sistem informasi pengolahan data simpan pinjam koperasi pegawai republik indonesia karya dharma kabupaten majene	RPL
2	Implementasi *Elimination And Choice Translation Reality* (*Electre)* Untuk Menentukan Calon Pembimbing Tugas Akhir Mahasiswa	implementasi elimination and choice translation reality *(electre)* untuk menentukan calon pembimbing tugas akhir mahasiswa	Sistem Cerdas
3	Sistem Otomasi Pada Smart Home Berbasis Internet Of Things (*Iot*)	sistem otomasi pada smart home berbasis *internet of things* (*Iot*)	*Iot*

3.3 Filtering/Stopword

Pada tahap ini, dilakukan penghapusan atau penyaringan kata-kata yang dapat mencerminkan isi dari suatu dokumen dengan menggunakan kamus *stopword* yang telah tersedia untuk menentukan kata-kata mana yang harus dihilangkan. Tabel 5 adalah hasil data yang telah dilakukan proses *filtering/Stopword.*

Tabel 5 Hasil *Filtering/Stopword* Data Judul Skripsi

No	Judul	Hasil Tokenisasi	Hasil Filtering	Kategori
1	Sistem Informasi Pengolahan Data Simpan Pinjam Koperasi Pegawai Republik Indonesia Karya Dharma Kabupaten Majene	Sistem Informasi Pengolahan Koperasi Data dan seterusnya	System informasi pengolahan koperasi data …	RPL
2	Implementasi Elimination and Choice Translation Reality (Electre) Untuk Menentukan Calon PembimbingTugas Akhir Mahasiswa	Calon Pembimbing Tugas Akhir Mahasiswa dan seterusnya	calon pembimbing tugas akhir mahasiswa …	Sistem Cerdas

		Otomasi		IoT
3	Sistem Otomasi Pada Smart Home Berbasis Internet Of Things (Iot)	Smart Internet of Things Home Jaringan Dan Seterusnya	Otomasi Smart Internet of Things Home Jaringan ...	

3.4 Stemming

Stemming adalah langkah dalam proses untuk mendapatkan kata dasar atau inti dari hasil *filtering*. Pada tahap ini, dilakukan penghapusan awalan dan akhiran kata untuk menghasilkan kata dasar dari setiap kata (token). Tabel 6 adalah hasil data yang telah dilakukan proses *Stemming*.

Tabel 7 Hasil *Stemming* Data Judul Skripsi

No	Judul	Hasil Tokenisasi	Hasil Filtering	Kategori
1	Sistem Informasi Pengolahan Data Simpan Pinjam Koperasi Pegawai Republik Indonesia Karya Dharma Kabupaten Majene	Informasi Pengolahan Data Simpan Pinjam dan seterusnya	informasi olah data Simpan Pinjam …	RPL
2	Implementasi *Elimination and Choice Translation Reality (Electre)* Untuk Menentukan Calon PembimbingTugas Akhir Mahasiswa	*elimination* *And* *Choice* *Reality* *Electre* dan seterusnya	*elimination* *and* *choice* *reality* *electre* …	SistemCerdas
3	Sistem Otomasi Pada Smart Home Berbasis Internet Of Things (Iot)	*Smart* *Home* Berbasis *Intermet* *Of* dan seterusnya	*smart* *home* basis *intermet* *of* …	IoT

3.5 Menentukan Atribut kata

Target kelas klasifikasi yang dipilih adalah RPL, Sistem Cerdas dan *Iot.* Selanjutnya dilakukan penentuan atribut kata (*Token).* Atribut kata ditentukan berdasrkan hasil hitung frekunsi kata atau kemunculan kata yang benar-benar merepresentasikan isi dari suatu judul skripsi tersebut. Karena semakin sering kata tersebut muncul maka dapat dikatakan bahwa kata tersebut mencirikan suatu kelas tertentu. 10 kata yang paling banyak muncul akan dipilih sebagai atribut kata. Atribut kata yang dipilih dapat dilihat pada tabel 8 berikut.

Tabel 8 Atribut Kata

RPL	Sistem Cerdas	IoT
Sistem	Implementasi	Otomatis
Perancangan	Algoritma	IoT
Informasi	Citra	Deteksi
Dan seterusnya …	Dan seterusnya …	Dan seterusnya …

(Sumber: Teknik Informatika Unsulbar, 2022)

3.6 Tahapan pembobotan (TF-IDF)

Pembobotan melibatkan proses mengubah kata menjadi representasi angka. TF merupakan frekuensi kemunculan *term* dalam dokumen yang relevan. Semakin tinggi jumlah atau nilai bobot suatu *term* dalam dokumen, semakin besar juga bobotnya dan memberikan nilai kesesuaian yang lebih tinggi. Sementara itu, TF-IDF adalah metode untuk menghitung nilai TF dan IDF untuk setiap token (kata) dalam setiap dokumen. Probabilitas kapasitas untuk menentukan kelas klasifikasi didasarkan pada kemunculan kata-kata tersebut.

3.7 Pengujian Sistem

Sistem menggunakan hasil klasifikasi untuk menampilkan tiga judul skripsi yang tidak sesuai dengan kategori kelasnya. Untuk mengetahui akurasi, jumlah data testing yang sesuai dengan target kelas yang telah ditentukan dibagi dengan jumlah keseluruhan data testing.untuk menghitung nilai akurasi dapat menggunakan persamaan 3 sehingga hasil akurasi dari uji coba dapat dilihat pada sebagai berikut:

$$Akurasi = \frac{97}{100} x\ 100\% = 97\%$$

Berdasarkan hasil pengolahan 150 data *training* dan 100 data *testing* menggunakan *naïve bayes* menunjukkan bahwa tingkat akurasi yang dihasilkan sebesar 97%. Untuk mencari nilai akurasi, *precision* dan *recall* dapat digunakan nilai *true positif, false negatif, true negatif*, dan *false positif.* Untuk mengetahui nilai-nilai tersebut dapat diketahui melalui tabel *confusion matriks.*

Tabel 9 Hasil *Confusion matrix*

Klasifikasi	TP	FP	TN	FN	*Accuracy*	*Precision*	*Recall*
RPL	32	1	66	1	0.98	0.97	0.97
SistemCerdas	31	1	66	2	0.97	0.969	0.939
IoT	34	1	65	0	0.99	0.971	1
Rata-rata					0.98	0.97	0.97

4. KESIMPULAN

Dari penelitian sebelumnya, dapat disimpulkan bahwa pengklasifikasian judul skripsi menggunakan algoritma *naive bayes* mampu mengelompokkan judul skripsi sesuai dengan topiknya. Dalam pengujian akurasi menggunakan 150 data *training* dan 100 data *testing*, diperoleh tingkat akurasi rata-rata sebesar 98%. Selain itu, presisi rata-rata sebesar 0,97% dan *Recall* rata-rata sebesar 0,97% juga didapatkan dari pengujian tersebut. Hasil pengujian menunjukkan bahwa implementasi metode *naive bayes* dalam klasifikasi judul skripsi telah berhasil, seperti yang dapat dilihat dari tingkat akurasi yang tinggi pada 100 data *testing* dan 150 data. Hal ini menunjukkan bahwa sistem yang dibangun telah memenuhi kebutuhan pengguna, yaitu dapat mengklasifikasikan judul berdasarkan kategori topik judul *naïve bayes.*

REFERENSI

[1] A. Rai, "What is Text Mining: Techniques and Applications," 6 Oktober 2022. [Online]. Available: https://www.upgrad.com/blog/what-is-text-mining-techniques-and-applications/.

[2] T. Syahdrajat, "Panduan Menulis Tinjauan Pustaka, Laporan Kasus Dan Artiikel Penelitian Di Juranl Kedokteran," *Dian Rakyat,* 2012.

[3] Barnawi, "Teknik Penulisan Karya Ilmiah," *Ar-Ruzz Media,* 2015.

[4] Mandias, G. F., "Penerapan data mining untuk evaluasi kinerja akademik mahasiswa di Universitas Klabat dengan metode klasifikasi," *Proceedings Konferensi Nasional Sistem dan Informatika (KNS&I),* 2015.

[5] R. Khalida dan H. Kusuma Bharata, "Analisa Komparasi Tiga Metode Data Mining dalam Prediksi Impor Komoditas Tanaman Biofarmaka," *Jurnal Ilmiah Komputasi,* Vol. %1 dari %2Vol. 19, No. 2, 2020.

[6] A. Firdaus dan W. Istalama Firdaus, "Text Mining Dan Pola Algoritma Dalam Penyelesaian Masalah Informasi : (Sebuah Ulasan)," *Jurnal JUPITER,* Vol. %1 dari %213, No. 1, pp. 66-78, 2021.

[7] I. Oktanisa dan A. Afif Supianto, "Perbandingan Teknik Klasifikasi Dalam Data Mining Untuk Bank Direct Marketing," *Jurnal Teknologi Informasi dan Ilmu Komputer,* Vol. %1 dari %25, No. 5 , 2018.

[8] S. Syarli, "Metode *Naive bayes* Untuk Prediksi Kelulusan (Studi Kasus: Data Mahasiswa Baru Perguruan Tinggi)," *Jurnal Ilmiah Ilmu Komputer,* Vol. %1 dari %22, No. 1, pp. 22-26, 2016.

[9] Rennie J, Shih L, Teevan J, and Karger D. Tackling, "The Poor Assumptions of *Naive bayes* Classifiers," *In Proceedings of the Twentieth International Conference on Machine Learning,* 2003.

[10] H. Febtadianrano Putro, R. Tri Vulandari dan W. Laksito Yuly Saptomo, "Penerapan Metode *Naive bayes* Untuk Klasifikasi Pelanggan," *Jurnal TIKomSin,* pp. 8, No. 2, 2020.

[11] Suyanto, "Machine Learning Tingkat Dasar dan Lanjut," *Bandung: Informatika Bandung.*, 2018.