

BAB III. METODOLOGI PENELITIAN

3.1. Metode Pengumpulan Data

Teknik pengumpulan data yang dilakukan oleh penulis yaitu:

3.1.1 Studi Literatur

Studi literatur merupakan sebuah metode pengumpulan data dengan membaca jurnal, buku yang dijadikan sebagai referensi yang berhubungan dengan penelitian ini.

3.1.2 Metode Kuantitatif

Metode kuantitatif digunakan untuk mengumpulkan *dataset* yang diperlukan dalam proses klasifikasi untuk menghasilkan data model klasifikasi serta mengumpulkan data judul publikasi dosen JTI POLINEMA. *Dataset* berisi kumpulan judul publikasi umum beserta label bidang penelitiannya yang diambil dari *website* Google Scholar.

Jumlah *dataset* yang dikumpulkan adalah 1900 data judul publikasi. Sedangkan jumlah data judul publikasi dosen JTI POLINEMA yang dikumpulkan menyesuaikan jumlah judul publikasi yang ditampilkan pada profil *Google Scholar* pada masing-masing dosen.

Dataset dikumpulkan melalui proses *crawling* dan diolah menggunakan metode *Naive Bayes* untuk menghasilkan data model klasifikasi *Naive Bayes*. Data model klasifikasi *Naive Bayes* yang dihasilkan digunakan sebagai dasar untuk melakukan klasifikasi publikasi dosen JTI POLINEMA.

Dataset terdiri dari 38 label bidang penelitian dan setiap label bidang penelitian terdiri dari 50 judul publikasi. *Dataset* akan dibagi menjadi data *training* dan data *testing*. *Data training* berisi kumpulan judul publikasi dan label bidang penelitiannya, sedangkan *Data testing* berisi kumpulan judul publikasi tanpa label bidang penelitiannya.

Rincian pembagian *dataset* yaitu, *data testing* sejumlah 380 data dan *data training* sejumlah 1520 data. *Data training* dan *data testing* tersebut digunakan untuk membuat sebuah model klasifikasi dengan menerapkan algoritma *Naive Bayes*. Model klasifikasi *Naive Bayes* yang dihasilkan berisi kumpulan judul publikasi beserta label bidang penelitiannya berdasarkan setiap grup riset JTI

POLINEMA. Pada tabel 3.1 ditampilkan bidang penelitian pada setiap grup riset JTI POLINEMA.

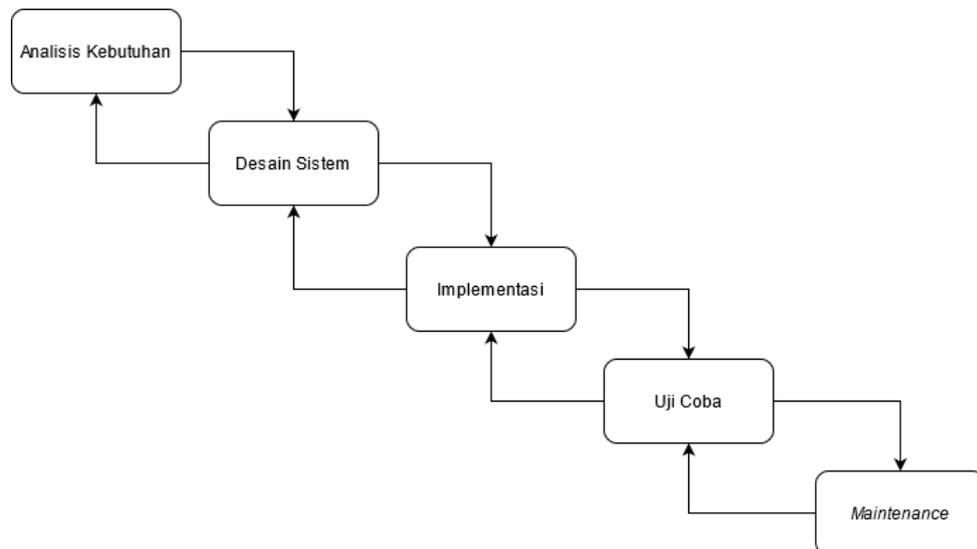
Tabel 3.1 Bidang Penelitian

No	Grup Riset	Bidang Penelitian
1	Sistem Cerdas	<i>Data Scalability and Analytics</i> <i>Artificial Intelligence</i> <i>Big Data</i> <i>Information Retrieval</i> <i>Expert System</i> <i>Data Mining</i> <i>Text Mining</i> <i>Business Intelligent</i> <i>Smart Living</i>
2	Aristektur Jaringan Komputer dan Data Security	<i>Cloud computing</i> IoT <i>Security Data</i> <i>Cyber Security</i> Virtualisasi <i>Network Technology</i> <i>Wireless Technology</i> <i>IT Forensic pada Jaringan</i> Enkripsi
3	<i>Computer Vision</i>	<i>Pattern Recognition</i> <i>Watermarking</i> Kompresi Pengolahan Citra Transformasi Filtering <i>Morphology</i>
4	<i>Information System</i>	<i>System Integration and Architecture</i> <i>Information Management</i>

		<i>Geographic Information System</i> Sistem Peramalan Sistem Pendukung Keputusan <i>E-Business</i> CRM <i>Technology Enhanced Learning (TEL)</i> <i>Computational Thinking</i>
5	Multimedia dan Game	<i>Game Development</i> Multimedia <i>Augmented Reality</i>

3.2. Metode Pengembangan Sistem

Dalam rancang bangun sistem informasi hasil klasifikasi bidang penelitian dosen JTI POLINEMA, penulis menggunakan model *waterfall*. Waterfall merupakan metode pengembangan perangkat lunak tradisional yang sistematis. Metode ini memiliki empat tahapan proses yang digunakan, diantaranya adalah analisis kebutuhan, *design system* (desain sistem), *implementation* (implementasi), dan *testing* (pengujian). Seperti yang ditampilkan pada Gambar 3.1.



Gambar 3.1 *Software Development Life Cycle (SDLC) Model Waterfall*.

3.2.1 Analisis Kebutuhan

Tahapan pertama adalah melakukan analisis kebutuhan terhadap sistem yang akan dibuat. Pada tahapan ini dilakukan identifikasi masalah terkait dengan aplikasi/sistem yang akan dibuat. Proses identifikasi tersebut dilakukan dengan menerima semua *input* berupa data yang diperlukan dalam membuat suatu sistem/aplikasi yang baik.

3.2.2 Desain Sistem

Desain sistem merupakan tahapan untuk melakukan perancangan suatu sistem/aplikasi yang akan diimplementasikan berdasarkan data yang diperoleh dari proses analisis kebutuhan. Rancangan yang dibuat berupa gambaran, perencanaan, sketsa dari sistem/aplikasi yang akan diimplementasikan.

3.2.3 Implementasi

Implementasi merupakan tahapan dimana hasil perancangan dari proses desain sistem diubah menjadi kode-kode program yang terintegrasi membentuk sebuah sistem/aplikasi yang lengkap.

3.2.4 Uji Coba

Uji Coba merupakan tahapan dimana aplikasi/sistem hasil implementasi dilakukan proses pengujian. Proses pengujian tersebut dilakukan untuk memastikan apakah sistem/aplikasi yang dibuat telah berfungsi dengan baik dan sesuai dengan kebutuhan yang telah dirancang sebelumnya.

3.2.5 Maintenance

Maintenance merupakan tahapan pemeliharaan terhadap sistem/aplikasi yang sudah melewati tahapan implementasi dan uji coba. Proses pemeliharaan tersebut berupa instalasi dan proses perbaikan sistem apabila ditemukan kesalahan/*bug* yang tidak ditemukan pada tahapan uji coba.

3.3 Metode Pengolahan Data

Teknik pengolahan data pada penelitian ini terbagi menjadi delapan tahapan untuk menghasilkan model klasifikasi *Naive Bayes* yang akan digunakan pada proses klasifikasi kecenderungan bidang penelitian dosen JTI POLINEMA. Delapan tahapan tersebut yaitu proses *crawling dataset* judul publikasi, *split dataset*, *preprocessing*, *term weighting* (TF-IDF), klasifikasi, perhitungan

confusion matrix, *crawling* judul publikasi dosen, klasifikasi judul publikasi dosen, dan penentuan kecenderungan bidang penelitian dosen.

3.3.1 Proses *Crawling Dataset*

Tahap pertama adalah melakukan proses *crawling* untuk mengumpulkan *dataset*. *Dataset* tersebut adalah daftar judul publikasi umum yang sudah memiliki label bidang penelitian. Daftar judul publikasi tersebut diambil melalui *website* Google Scholar.

Jumlah *dataset* yang dikumpulkan adalah 1900 data. *Dataset* tersebut dikumpulkan melalui proses *crawling* dan diolah menggunakan metode *Naive Bayes* untuk menghasilkan data model klasifikasi *Naive Bayes*. Data model klasifikasi *Naive Bayes* yang dihasilkan digunakan sebagai dasar untuk melakukan klasifikasi publikasi dosen JTI POLINEMA.

Berikut adalah tahapan dari proses *crawling* yang dilakukan:

1. Memasukkan *Link Google Scholar* yang didalamnya terdapat kata kunci yaitu berupa bidang penelitian untuk setiap grup riset JTI POLINEMA serta nomor halaman daftar hasil pencarian ke dalam URL. Bidang penelitian yang dimasukkan sebagai kata kunci pada *Link Google Scholar* menggunakan Bahasa Indonesia
2. Melakukan *request* terhadap URL yang berisi *Link Google Scholar* dengan menggunakan *User Agent* yang dimiliki *web browser*.
3. Melakukan proses menampilkan halaman *Google Scholar* yang berisi daftar judul publikasi berbahasa Indonesia menggunakan *Library BeautifulSoup*. Dengan menggunakan hasil dari proses *request* yang sudah dilakukan.
4. Melakukan pencarian terhadap judul publikasi berbahasa Indonesia pada halaman *Google Scholar* yang sudah ditampilkan menggunakan *Library BeautifulSoup*.
5. Memasukkan judul publikasi ke dalam file berekstensi *.csv* untuk dilakukan proses klasifikasi. Di dalam file tersebut terdapat daftar judul publikasi beserta label bidang penelitiannya.

3.3.2 Proses *Split Dataset*

Setelah dilakukan proses *crawling* selanjutnya adalah melakukan tahap *split dataset*. Yaitu membagi dataset menjadi *data training* dan *data testing*. *Data training* adalah kumpulan data judul publikasi yang memiliki label bidang penelitian. Sedangkan *data testing* adalah kumpulan data judul publikasi tanpa label bidang penelitian.

Proses *split dataset* menggunakan metode *K-Fold Cross Validation* dengan nilai $k = 5$, *dataset* dibagi menjadi lima subset. Jumlah *dataset* yang digunakan adalah 1900 data, maka satu subset *dataset* berjumlah 380 data. Sehingga dilakukan iterasi untuk melakukan proses *split dataset* sejumlah lima kali perulangan.

Pada Tabel 3.2 ditampilkan pembagian *data training* dan *data testing* pada *dataset* untuk seluruh kategori bidang penelitian secara rekursif sebanyak lima kali perulangan.

Tabel 3.2 Pembagian *Dataset* Seluruh Kategori Bidang Penelitian

Iterasi	Jumlah <i>Data Training</i> dan <i>Data Testing</i>				
	1	380 (<i>data testing</i>)	380 (<i>data training</i>)	380 (<i>data training</i>)	380 (<i>data training</i>)
2	380 (<i>data training</i>)	380 (<i>data testing</i>)	380 (<i>data training</i>)	380 (<i>data training</i>)	380 (<i>data training</i>)
3	380 (<i>data training</i>)	380 (<i>data training</i>)	380 (<i>data testing</i>)	380 (<i>data training</i>)	380 (<i>data training</i>)
4	380 (<i>data training</i>)	380 (<i>data training</i>)	380 (<i>data training</i>)	380 (<i>data testing</i>)	380 (<i>data training</i>)
5	380 (<i>data training</i>)	380 (<i>data training</i>)	380 (<i>data training</i>)	380 (<i>data training</i>)	380 (<i>data testing</i>)

Pada Tabel 3.3 ditampilkan pembagian *data training* dan *data testing* pada *dataset* salah satu kategori bidang penelitian. Kategori yang digunakan terdiri dari 38 bidang penelitian sehingga satu kategori bidang penelitian terdiri dari 50 *dataset*.

Tabel 3.3 Pembagian *Dataset* Salah Satu Kategori Bidang Penelitian

Iterasi	Jumlah <i>Data Training</i> dan <i>Data Testing</i>				
	1	10 (<i>data testing</i>)	10 (<i>data training</i>)	10 (<i>data training</i>)	10 (<i>data training</i>)
2	10 (<i>data training</i>)	10 (<i>data testing</i>)	10 (<i>data training</i>)	10 (<i>data training</i>)	10 (<i>data training</i>)
3	10 (<i>data training</i>)	10 (<i>data training</i>)	10 (<i>data testing</i>)	10 (<i>data training</i>)	10 (<i>data training</i>)
4	10 (<i>data training</i>)	10 (<i>data training</i>)	10 (<i>data training</i>)	10 (<i>data testing</i>)	10 (<i>data training</i>)
5	10 (<i>data training</i>)	10 (<i>data training</i>)	10 (<i>data training</i>)	10 (<i>data training</i>)	10 (<i>data testing</i>)

Pada Tabel 3.4 adalah contoh dokumen yang berisi judul publikasi yang memiliki label bidang penelitian digunakan sebagai *data training*.

Tabel 3.4 Contoh *Data Training*

No	Label Dokumen	Judul Publikasi
1.	Text Mining (D1)	Penggunaan fitur worldcloud dan document term matrix dalam text mining
2.	Artificial Intelligence (D2)	Estimasi indeks kualitas air menggunakan pendekatan artificial intelligence dan multi-linear regression

Pada Tabel 3.5 adalah contoh dokumen yang berisi judul publikasi yang belum memiliki label bidang penelitian digunakan sebagai *data testing*.

Tabel 3.5 Contoh *Data Testing*

No	Judul Publikasi
1.	Menerapkan Artificial Intelligence dalam Fitur Worldcloud pada Document
2.	Menerapkan Multi-linear Regression pada Index Matrix

3.3.3 *Preprocessing*

Setelah data dibagi menjadi *data training* dan *data testing* selanjutnya dilakukan *preprocessing* pada kedua data tersebut. Pada tahap *preprocessing* pertama dilakukan proses pengubahan seluruh huruf pada kata menjadi huruf kecil

(*Case Folding*). Kemudian membuang semua kata-kata yang bersifat umum (*Stopword Removal*). Dan melakukan proses perubahan kata pada suatu dokumen menjadi kata dasar (*Stemming*). Terakhir dilakukan proses pemenggalan kata pada setiap dokumen (*Tokenization*). Pada Tabel 3.6 ditampilkan contoh hasil dari *preprocessing* untuk *data training*.

Tabel 3.6 Contoh Hasil *Preprocessing Data Training*

No	Label dokumen	Judul Publikasi	Term Hasil <i>Preprocessing</i>
1.	Text Mining (D1)	Penggunaan fitur worldcloud dan document term matrix dalam text mining	fitur, worldcloud, document, term, matrix, text, mining
2.	Artificial Intelligence (D2)	Estimasi indeks kualitas air menggunakan pendekatan artificial intelligence dan multi-linear regression	estimasi, indeks, artificial, intelligence, multi-linear, regression

Pada Tabel 3.7 ditampilkan contoh hasil dari *preprocessing* untuk *data testing*.

Tabel 3.7 Contoh Hasil *Preprocessing Data Testing*

No	Judul Publikasi	Term Judul Publikasi	Term Hasil <i>Preprocessing</i>
1.	Menerapkan Artificial Intelligence dalam Fitur Worldcloud pada Document	menerapkan, artificial, intelligence, dalam, fitur, worldcloud, pada, document	artificial, intelligence, fitur, worldcloud, document
2.	Menerapkan Multi-linear Regression pada Index	menerapkan, multi-linear,	Multi-linear, regression, index, matrix

	Matrix	regression, pada, index, matrix	
--	--------	------------------------------------	--

3.3.4 Term Weighting

Setelah dilakukan *preprocessing* pada data *training* dan data *testing*. Selanjutnya dilakukan proses *term weighting* untuk setiap kata pada dokumen di data *training* menggunakan metode TF-IDF. Berikut adalah contoh perhitungan metode TF-IDF:

1. Menghitung jumlah kemunculan kata pada setiap dokumen $TF(d, t)$. Hasil perhitungan ditampilkan pada Tabel 3.8

Tabel 3.8 Hasil Perhitungan $TF(d, t)$

No	Term	$TF(1, t)$	$TF(2, t)$
1.	fitur	$TF(1, 1) = 1$	$TF(2, 1) = 0$
2.	worldcloud	$TF(1, 2) = 1$	$TF(2, 2) = 0$
3.	document	$TF(1, 3) = 1$	$TF(2, 3) = 0$
4.	term	$TF(1, 4) = 1$	$TF(2, 4) = 0$
5.	matrix	$TF(1, 5) = 1$	$TF(2, 5) = 0$
6.	text	$TF(1, 6) = 1$	$TF(2, 6) = 0$
7.	mining	$TF(1, 7) = 1$	$TF(2, 7) = 0$
8.	estimasi	$TF(1, 8) = 0$	$TF(2, 8) = 1$
9.	indeks	$TF(1, 9) = 0$	$TF(2, 9) = 1$
10.	artificial	$TF(1, 10) = 0$	$TF(2, 10) = 1$
11.	intelligence	$TF(1, 11) = 0$	$TF(2, 11) = 1$
12.	multi-linear	$TF(1, 12) = 0$	$TF(2, 12) = 1$
13.	regression	$TF(1, 13) = 0$	$TF(2, 13) = 1$

2. Melakukan perhitungan *document frequency* $df(t)$ yaitu banyaknya dokumen dimana suatu *term* t muncul. Hasil perhitungan ditampilkan pada Tabel 3.9.

Tabel 3.9 Hasil Perhitungan $df(t)$

No	Term	$df(t)$
1.	fitur	$df(1) = 1$
2.	worldcloud	$df(2) = 1$
3.	document	$df(3) = 1$
4.	term	$df(4) = 1$
5.	matrix	$df(5) = 1$
6.	text	$df(6) = 1$
7.	mining	$df(7) = 1$
8.	estimasi	$df(8) = 1$
9.	indeks	$df(9) = 1$
10.	artificial	$df(10) = 1$
11.	intelligence	$df(11) = 1$
12.	multi-linear	$df(12) = 1$
13.	regression	$df(13) = 1$

3. Menghitung *invers document frequency* $IDF(t)$. Jumlah dokumen (N) adalah 2. Hasil perhitungan ditampilkan pada Tabel 3.10

Tabel 3.10 Hasil Perhitungan $IDF(t)$

No	Term	$IDF(t)$
1.	fitur	$IDF(1) = 1 + \log\left(\frac{Nd}{df(1)}\right) = 1.30103$
2.	worldcloud	$IDF(t) = 1 + \log\left(\frac{Nd}{df(2)}\right) = 1.30103$
3.	document	$IDF(t) = 1 + \log\left(\frac{Nd}{df(3)}\right) = 1.30103$

4.	term	$IDF(t) = 1 + \log\left(\frac{Nd}{df(4)}\right) = 1.30103$
5.	matrix	$IDF(t) = 1 + \log\left(\frac{Nd}{df(5)}\right) = 1.30103$
6.	text	$IDF(t) = 1 + \log\left(\frac{Nd}{df(6)}\right) = 1.30103$
7.	mining	$IDF(t) = 1 + \log\left(\frac{Nd}{df(7)}\right) = 1.30103$
8.	estimasi	$IDF(t) = 1 + \log\left(\frac{Nd}{df(8)}\right) = 1.30103$
9.	indeks	$IDF(t) = 1 + \log\left(\frac{Nd}{df(9)}\right) = 1.30103$
10.	artificial	$IDF(t) = 1 + \log\left(\frac{Nd}{df(10)}\right) = 1.30103$
11.	intelligence	$IDF(t) = 1 + \log\left(\frac{Nd}{df(11)}\right) = 1.30103$
12.	multi-linear	$IDF(t) = 1 + \log\left(\frac{Nd}{df(12)}\right) = 1.30103$
13.	regression	$IDF(t) = 1 + \log\left(\frac{Nd}{df(13)}\right) = 1.30103$

4. Melakukan perhitungan TF.idf. Pada Tabel 3.11 ditampilkan hasil dari perhitungan tersebut.

Tabel 3.11 Hasil Perhitungan TF.Idf

No	Term	TF.Idf	
		TF(1, t).IDF(t)	TF(2, t).IDF(t)
1.	fitur	$W_{1,1} = TF(1, 1).IDF(1) = 1.30103$	$W_{2,1} = TF(2, 1).IDF(1) = 0$
2.	world-cloud	$W_{1,2} = TF(1, 2).IDF(2)$	$W_{2,2} = TF(2, 2).IDF(2) = 0$

		= 1.30103	
3.	document	$W_{1,3} = TF(1, 3).IDF(3)$ = 1.30103	$W_{2,3} = TF(2, 3).IDF(3) = 0$
4.	term	$W_{1,4} = TF(1, 4).IDF(4)$ = 1.30103	$W_{2,4} = TF(2, 4).IDF(4) = 0$
5.	matrix	$W_{1,5} = TF(1, 5).IDF(5)$ = 1.30103	$W_{2,5} = TF(2, 5).IDF(5) = 0$
6.	text	$W_{1,6} = TF(1, 6).IDF(6)$ = 1.30103	$W_{2,6} = TF(2, 6).IDF(6) = 0$
7.	Mining	$W_{1,7} = TF(1, 7).IDF(7)$ = 1.30103	$W_{2,7} = TF(2, 7).IDF(7) = 0$
8.	estimasi	$W_{1,8} = TF(1, 8).IDF(8)$ = 0	$W_{2,8} = TF(2, 8).IDF(8)$ = 1.30103
9.	indeks	$W_{1,9} = TF(1, 9).IDF(9)$ = 0	$W_{2,9} = TF(2, 9).IDF(9)$ = 1.30103
10.	artificial	$W_{1,10} = TF(1, 10).IDF(10)$ = 0	$W_{2,10} = TF(2, 10).IDF(10)$ = 1.30103
11.	intelligence	$W_{1,11} = TF(1, 11).IDF(11)$ = 0	$W_{2,11} = TF(2, 11).IDF(11)$ = 1.30103
12.	multi-linear	$W_{1,12} = TF(1, 12).IDF(12)$ = 0	$W_{2,12} = TF(2, 12).IDF(12)$ = 1.30103
13.	regression	$W_{1,13} = TF(1, 13).IDF(13)$ = 0	$W_{2,13} = TF(2, 13).IDF(13)$ = 1.30103

3.3.5 Klasifikasi *Naive Bayes*

Selanjutnya dilakukan proses klasifikasi menggunakan data *training* yang sudah dilakukan proses *term weighting* dan data *testing*. Data *training* dan data

testing tersebut sudah dilakukan *preprocessing*. Proses klasifikasi menerapkan metode *Naive Bayes*. Berikut adalah contoh dari perhitungan dengan metode *Naive Bayes*:

1. Menghitung probabilitas kategori dokumen $p(c_j)$ pada data *training*.

Perhitungan ditampilkan pada persamaan 3.1 dan 3.2

$$p(c_1) = \frac{1}{2} \quad (3.1)$$

$$p(c_2) = \frac{1}{2} \quad (3.2)$$

2. Menghitung probabilitas kata pada setiap kategori $p(t_i|c_j)$ pada data *training*. Jumlah kata pada masing-masing kategori sejumlah $n = 13$. Pada Tabel 3.12 ditampilkan hasil dari perhitungan tersebut.

Tabel 3.12 Hasil Perhitungan $p(t_i|c_j)$

No	Term	Label Dokumen	
		Text Mining	Artificial Intelligence
1.	Fitur	$p(t_1 c_1) = \left(\frac{W_{1,1}+1}{\sum_{i=1}^n Wi+ kosakata } \right) = 0.104085047$	$p(t_1 c_2) = \left(\frac{W_{1,2}+1}{\sum_{i=1}^n Wi+ kosakata } \right) = 0.048062643$
2.	Worldcloud	$p(t_2 c_1) = \left(\frac{W_{2,1}+1}{\sum_{i=1}^n Wi+ kosakata } \right) = 0.104085047$	$p(t_2 c_2) = \left(\frac{W_{2,2}+1}{\sum_{i=1}^n Wi+ kosakata } \right) = 0.048062643$
3.	document	$p(t_3 c_1) = \left(\frac{W_{3,1}+1}{\sum_{i=1}^n Wi+ kosakata } \right) = 0.104085047$	$p(t_3 c_2) = \left(\frac{W_{3,2}+1}{\sum_{i=1}^n Wi+ kosakata } \right) = 0.048062643$
4.	term	$p(t_4 c_1) = \left(\frac{W_{4,1}+1}{\sum_{i=1}^n Wi+ kosakata } \right) = 0.104085047$	$p(t_4 c_2) = \left(\frac{W_{4,2}+1}{\sum_{i=1}^n Wi+ kosakata } \right) = 0.048062643$
5.	matrix	$p(t_5 c_1) = \left(\frac{W_{5,1}+1}{\sum_{i=1}^n Wi+ kosakata } \right) = 0.104085047$	$p(t_5 c_2) = \left(\frac{W_{5,2}+1}{\sum_{i=1}^n Wi+ kosakata } \right) = 0.048062643$
6.	text	$p(t_6 c_1) = \left(\frac{W_{6,1}+1}{\sum_{i=1}^n Wi+ kosakata } \right) =$	

		0.104085047	$p(t_6 c_2) = \left(\frac{W_{6,2}+1}{\sum_{i=1}^n W_i+ kosakata } \right) =$ 0.048062643
7.	mining	$p(t_7 c_1) = \left(\frac{W_{7,1}+1}{\sum_{i=1}^n W_i+ kosakata } \right) =$ 0.104085047	$p(t_7 c_2) = \left(\frac{W_{7,2}+1}{\sum_{i=1}^n W_i+ kosakata } \right) =$ 0.048062643
8.	estimasi	$p(t_8 c_1) = \left(\frac{W_{8,1}+1}{\sum_{i=1}^n W_i+ kosakata } \right) =$ 0.045234111	$p(t_8 c_2) = \left(\frac{W_{8,2}+1}{\sum_{i=1}^n W_i+ kosakata } \right) =$ 0.110593583
9.	indeks	$p(t_9 c_1) = \left(\frac{W_{9,1}+1}{\sum_{i=1}^n W_i+ kosakata } \right) =$ 0.045234111	$p(t_9 c_2) = \left(\frac{W_{9,2}+1}{\sum_{i=1}^n W_i+ kosakata } \right) =$ 0.110593583
10.	artificial	$p(t_{10} c_1) = \left(\frac{W_{10,1}+1}{\sum_{i=1}^n W_i+ kosakata } \right) =$ 0.045234111	$p(t_{10} c_2) = \left(\frac{W_{10,2}+1}{\sum_{i=1}^n W_i+ kosakata } \right) =$ 0.110593583
11.	intelligence	$p(t_{11} c_1) = \left(\frac{W_{11,1}+1}{\sum_{i=1}^n W_i+ kosakata } \right) =$ 0.045234111	$p(t_{11} c_2) = \left(\frac{W_{11,2}+1}{\sum_{i=1}^n W_i+ kosakata } \right) =$ 0.110593583
12.	multi-linear	$p(t_{12} c_1) = \left(\frac{W_{12,1}+1}{\sum_{i=1}^n W_i+ kosakata } \right) =$ 0.045234111	$p(t_{12} c_2) = \left(\frac{W_{12,2}+1}{\sum_{i=1}^n W_i+ kosakata } \right) =$ 0.110593583
13.	regression	$p(t_{13} c_1) = \left(\frac{W_{13,1}+1}{\sum_{i=1}^n W_i+ kosakata } \right) =$ 0.045234111	$p(t_{13} c_2) = \left(\frac{W_{13,2}+1}{\sum_{i=1}^n W_i+ kosakata } \right) =$ 0.110593583

3. Membandingkan dan mencari nilai probabilitas terbesar c_{MAP} untuk menentukan kategori bidang penelitian pada *data testing*.

Sebelum mencari probabilitas terbesar c_{MAP} , dilakukan perhitungan untuk mencari nilai probabilitas untuk setiap kategori. Untuk hasil perhitungan pada data *testing* pada judul satu adalah sebagai berikut. Persamaan 3.3 adalah perhitungan pada kategori Text Mining, dan persamaan 3.4 adalah perhitungan pada kategori Artificial Intelligence.

$$\begin{aligned}
& p(c_1) \times p(t_1|c_1) \times p(t_2|c_1) \times p(t_3|c_1) \times p(t_{10}|c_1) \times \\
& p(t_{11}|c_1) \times p(t_{12}|c_1) \\
& = 5.21835E-08
\end{aligned} \tag{3.3}$$

$$\begin{aligned}
& p(c_1) \times p(t_1|c_2) \times p(t_2|c_2) \times p(t_3|c_2) \times p(t_{10}|c_2) \times \\
& p(t_{11}|c_2) \times p(t_{12}|c_2) \\
& = 7.50901E-08
\end{aligned} \tag{3.4}$$

Sehingga judul satu pada *data testing* termasuk dalam kategori Artificial Intelligence.

Untuk hasil perhitungan pada data *testing* pada judul dua adalah sebagai berikut. Persamaan 3.5 adalah perhitungan pada kategori Text Mining, dan persamaan 3.6 adalah perhitungan pada kategori Artificial Intelligence.

$$\begin{aligned}
& p(c_1) \times p(t_{12}|c_1) \times p(t_{13}|c_1) \times p(t_9|c_1) \times p(t_5|c_1) \\
& = 4.81678E-06
\end{aligned} \tag{3.5}$$

$$\begin{aligned}
& p(c_1) \times p(t_{12}|c_2) \times p(t_{13}|c_2) \times p(t_9|c_2) \times p(t_5|c_2) \\
& = 3.25063E-05
\end{aligned} \tag{3.6}$$

Sehingga judul dua pada *data testing* termasuk dalam kategori Text Mining.

Proses klasifikasi *Naive Bayes* menghasilkan data berupa model klasifikasi *Naive Bayes*. Model klasifikasi *Naive Bayes* selanjutnya digunakan sebagai dasar untuk melakukan proses klasifikasi judul publikasi dosen JTI POLINEMA. Hasil dari proses klasifikasi judul publikasi dosen digunakan untuk menentukan kecenderungan bidang penelitian dosen JTI POLINEMA.

3.3.6 Confusion Matrix

Model dari hasil klasifikasi menggunakan metode *Naive Bayes* selanjutnya di evaluasi menggunakan *confusion matrix*. Evaluasi menggunakan matriks ditampilkan pada Tabel 3.13.

Tabel 3.13 *Confusion Matrix*

Klasifikasi Sebenarnya	Diklasifikasikan Menjadi	
	+	-
+	True Positive	False Positive
-	False Negative	True Negative

Contoh hasil klasifikasi *Naive Bayes* ditampilkan pada tabel 3.14

Tabel 3.14 Hasil Klasifikasi Naive Bayes

No. Dokumen	Klasifikasi Sebenarnya	Hasil Klasifikasi	Keterangan
1.	Artificial Intelligence	Artificial Intelligence	TRUE
2.	Text Mining	Text Mining	TRUE

Evaluasi menggunakan matriks yang ditampilkan pada Tabel 3.15.

Tabel 3.15 Evaluasi *Confusion Matrix*

No.	Klasifikasi Dengan Naive Bayes	Klasifikasi Sebenarnya	
		Artificial Intelligence	Text Mining
1.	Artificial Intelligence	1	0
2.	Text Mining	0	1

Hasil Contoh dari perhitungan *precision*, akurasi, dan *recall* ditampilkan pada persamaan 3.7, 3.8, 3.9.

$$Precision = \frac{TP}{TP+FP} \times 100\% = 100\% \quad (3.7)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% = 100\% \quad (3.8)$$

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% = 100\% \quad (3.9)$$

3.3.7 *Crawling* Judul Publikasi Dosen

Proses *crawling* judul publikasi dosen bertujuan untuk mengumpulkan data judul publikasi setiap dosen JTI POLINEMA. Data judul publikasi dosen yang dikumpulkan selanjutnya akan dilakukan proses klasifikasi. Berikut adalah tahapan dari proses *crawling* judul publikasi dosen yang dilakukan:

1. Memasukkan *Link Google Scholar* profil dosen JTI POLINEMA.
2. Melakukan *request* terhadap URL yang berisi *Link Google Scholar* dengan menggunakan *User Agent* yang dimiliki *web browser*.
3. Melakukan proses menampilkan halaman *Google Scholar* yang berisi daftar judul publikasi yang dimiliki dosen menggunakan *Library Beautiful Soup*. Dengan menggunakan hasil dari proses *request* yang sudah dilakukan.
4. Melakukan pencarian terhadap judul publikasi dosen berbahasa Indonesia pada halaman *Google Scholar* yang sudah ditampilkan menggunakan *Library Beautiful Soup*. Proses pencarian judul publikasi berbahasa Indonesia dilakukan dengan cara melakukan pengecekan kata Bahasa Indonesia pada judul publikasi. Jika judul publikasi memiliki kata berbahasa Indonesia maka akan dikumpulkan ke dalam file berekstensi *.csv*.
5. Memasukkan judul publikasi dosen ke dalam file berekstensi *.csv* untuk dilakukan proses klasifikasi. Di dalam file tersebut terdapat daftar judul publikasi setiap dosen JTI POLINEMA beserta label bidang penelitiannya.

3.3.8 Klasifikasi Judul Publikasi Dosen

Proses klasifikasi judul publikasi dosen bertujuan untuk memberi label pada judul publikasi setiap dosen JTI POLINEMA. Hasil dari proses klasifikasi digunakan sebagai dasar untuk menentukan kecenderungan bidang penelitian

setiap dosen JTI POLINEMA. Berikut adalah tahapan dari proses klasifikasi judul publikasi dosen yang dilakukan:

1. Menghitung nilai probabilitas kata $p(t_i|c_j)$ pada data model klasifikasi *Naive Bayes*.
2. Menghitung nilai probabilitas setiap kategori bidang penelitian pada masing-masing judul publikasi di data judul publikasi dosen.
3. Mencari nilai probabilitas terbesar c_{MAP} pada masing-masing kategori bidang penelitian pada judul publikasi. Nilai probabilitas terbesar yang dimiliki oleh salah satu kategori digunakan untuk menentukan kategori bidang penelitian setiap judul publikasi di data judul publikasi dosen.
4. Memasukkan data judul publikasi dosen hasil klasifikasi ke dalam file berekstensi .csv

3.3.9 Penentuan Kecenderungan Bidang Penelitian Dosen

Proses penentuan kecenderungan bidang penelitian dosen bertujuan untuk menentukan kecenderungan bidang penelitian dan grup riset masing-masing dosen JTI POLINEMA berdasarkan pada hasil proses klasifikasi judul publikasi dosen. Berikut adalah tahapan dari proses penentuan kecenderungan bidang penelitian dosen yang dilakukan:

1. Menghitung jumlah setiap label bidang penelitian dan grup riset berdasarkan judul publikasi masing-masing dosen pada data judul publikasi dosen.
2. Mencari nilai tertinggi dari jumlah setiap label bidang penelitian dan grup riset berdasarkan judul publikasi masing-masing dosen pada data judul publikasi dosen.
3. Memasukkan data kecenderungan grup riset dan bidang penelitian dosen ke dalam file berekstensi .csv.