

Klasifikasi Berita Menggunakan Metode *Naïve Bayes Classifier*

Sayyid Muhammad Habib¹, Elin Haerani², Siska Kurnia Gusti³, Siti Ramadhani⁴

^{1,2,3,4}Teknik Informatika UIN Sultan Syarif Kasim Riau

Jl. H.R. Soebrantas No. 155 KM. 18 Simpang Baru, Pekanbaru 28293

Corresponding author's e-mail: 11850111414@students.uin-suska.ac.id¹, elin.haerani@uin-suska.ac.id²,

siskakurniagusti@uin-suska.ac.id³, siti.ramadhani@uin-suska.ac.id⁴

Abstrak - Tingginya kecenderungan masyarakat dalam mengakses berita secara online, membuat editor dan portal berita harus menyediakan berita yang berkualitas. Namun berita pada portal tersebut masih diklasifikasikan secara umum, sehingga ketika pembaca ingin mendapatkan kategori berita yang lebih spesifik harus dilakukan secara manual dengan menyaring berita-berita tersebut. Hal ini juga yang dialami oleh bidang sosial Badan Pusat Statistik Provinsi Riau yang kesulitan dalam mencari dan mengklasifikasikan jenis berita tentang Provinsi Riau. Oleh sebab itu, proses pengklasifikasian berita menggunakan metode *naïve bayes classifier* merupakan hal yang penting untuk dilakukan. Jumlah berita yang digunakan dalam penelitian ini berjumlah 510 berita dan dikategorikan menjadi 3 kategori yaitu demokrasi, kemiskinan, dan ketenagakerjaan. Agar membantu bidang sosial di Badan Pusat Statistik Provinsi Riau dalam mengklasifikasikan jenis berita sebagai landasan fenomena yang terjadi di daerah Provinsi Riau berdasarkan dari nilai indeks demokrasi, ketenagakerjaan, dan kemiskinan Provinsi Riau. Proses pengklasifikasian berita dalam penelitian ini meliputi: pengumpulan data, *text preprocessing*, pembobotan kata, dan klasifikasi *naïve bayes classifier*. Nilai akurasi tertinggi yang diperoleh dalam penelitian ini sebesar 94% dengan pembagian data uji 10% dan data latih 90%. Kata kunci: *Berita, Badan Pusat Statistik, Klasifikasi, Naïve Bayes Classifier, Riau*

Abstract - The high tendency of people to access news, especially online news, makes editors and news portal sites to provide quality information and news. However, the news grouping is still classified in general, so, when the reader want to get a more specific category of news, it must be done manually by filtering the news. This is also happened by the social sector of the Badan Pusat Statistik Provinsi Riau, which has difficulty in finding news about Riau Province. Therefore, the process of classifying news using the Naive Bayes Classifier method is an important thing to do. The number of news used in this research is 510 news and it's categorized into 3 categories, namely democracy, poverty, and employment. The news classification process in this research includes: data collection, manual labeling, text preprocessing, term weighting, and naive Bayes classifier classification. The highest accuracy value obtained in this research was 94% with the distribution of 10% test data and 90% training data.

Keywords: *Badan Pusat Statistik, Classification, Naïve Bayes Classifier, News, Riau*

1. Pendahuluan

Pada zaman digital saat ini, berita di media sosial lebih diminati dan menjadi sarana bagi masyarakat untuk mengetahui tentang informasi yang telah terjadi. Situs portal berita menjadi salah satu sumber informasi berita yang menyajikan dan mengumpulkan berita-berita dari berbagai sumber untuk ditampilkan kepada pembaca. Tingginya kecenderungan masyarakat dalam mengakses berita khususnya berita secara online, membuat editor dan situs portal berita bekerja lebih keras dalam memenuhi kebutuhan masyarakat untuk menyediakan informasi dan berita yang berkualitas. Pada portal berita biasanya terdapat kategori dari berita yang telah dikelompokkan dengan tujuan untuk mempermudah pembaca mencari berita yang diinginkan secara cepat. Namun pengelompokkan berita tersebut masih diklasifikasikan secara umum, sehingga ketika pembaca ingin mendapatkan kategori berita yang lebih spesifik lagi harus dilakukan secara manual dengan menyaring berita dari kategori tersebut dan mengklasifikasikannya menjadi subkategori yang lebih detail. Hal tersebut menjadi sulit karena harus membaca dan menyaring berita sedangkan jumlah berita terus meningkat setiap waktu dengan cepat, dan setiap berita memiliki tingkat similaritas yang tinggi. Oleh sebab itu, proses pengklasifikasian berita merupakan hal yang penting untuk dilakukan.

Proses klasifikasi termasuk ke dalam salah satu bidang *data mining*, yaitu pada bidang *text mining*. Proses pengklasifikasian mengarah kepada aktifitas mempelajari dan menganalisa dokumen teks yang *preclassified* untuk menghasilkan sebuah model yang dapat digunakan dalam proses pengelompokkan dokumen teks baru yang belum memiliki kelas. Klasifikasi merupakan proses untuk menetapkan suatu dokumen ke dalam sebuah kategori atau kelas yang sesuai dengan karakteristik dari dokumen tersebut [1]. Pada umumnya pemberian label pada proses klasifikasi masih dilakukan dengan cara manual oleh tim ahli pada *dataset* dalam jumlah data yang cukup banyak. Oleh sebab itu, proses pengklasifikasian akan dibuat dalam penelitian ini yang digunakan dalam proses pelabelan otomatis berskala besar untuk menggantikan proses pelabelan manual.

Metode *naïve bayes classifier* adalah salah satu metode yang memiliki akurasi cukup baik dalam proses klasifikasi. Metode *Naïve Bayes Classifier* merupakan sebuah metode yang mengaplikasikan Teorema Bayes dalam proses klasifikasi dengan probabilitas yang sederhana dan mengasumsikan ketidakketergantungan (independen) yang tinggi [2]. Metode *Naïve Bayes Classifier* terdapat beberapa keunikan, seperti perhitungan yang cepat, sederhana, dan berakurasi tinggi [3]. Pada jumlah data latih atau data *training* yang diproses pada metode *naïve bayes classifier* berjumlah kecil sebagai dasar untuk menghitung nilai estimasi parameter pada proses klasifikasi data. Hal tersebut dikarenakan variabel digunakan diasumsikan sebagai independent, sehingga hanya dibutuhkan sebuah varian dari suatu variabel pada kategori tertentu dalam penentuan klasifikasi dan bukan menggunakan keseluruhan dari matrix kovarians.

Penelitian yang akan dilakukan untuk mengklasifikasikan berita menggunakan metode *Naïve Bayes Classifier* untuk memperoleh tingkat akurasi yang baik dan pola klasifikasi berita yang menarik, sehingga mendapatkan hasil klasifikasi berita sesuai dengan kategori yang diinginkan. Adapun penelitian ini didasari oleh kesulitan yang dialami bidang sosial di Badan Pusat Statistik Provinsi Riau dalam mencari dan mengklasifikasikan jenis berita khususnya yang terjadi di daerah Provinsi Riau menjadi 3 kategori yaitu demokrasi, ketenagakerjaan, dan kemiskinan. Proses pencarian dan pengelompokkan berita tersebut juga masih dilakukan secara manual dengan mencari berita satu persatu pada portal berita dan media koran, kemudian dilakukan pelabelan manual pada berita tersebut yang jumlahnya cukup banyak. Sehingga proses pengklasifikasian berita tidak mendapatkan hasil yang optimal, karena dalam proses pelabelan berita masih dilakukan secara manual dengan harus membaca isi berita tersebut satupersatu, yang membutuhkan waktu serta tenaga yang cukup besar. Berita yang diklasifikasikan tersebut digunakan sebagai landasan fenomena dari nilai indeks demokrasi, nilai indeks ketenagakerjaan, dan nilai indeks kemiskinan Provinsi Riau.

Oleh karena itu, penelitian ini bertujuan untuk mengklasifikasikan jenis berita berdasarkan kategorinya dan memperoleh hasil akurasi klasifikasi menggunakan metode *naïve bayes classifier* yang terbaik, agar membantu bidang sosial di Badan Pusat Statistik Provinsi Riau dalam mencari dan pengelompokkan berita sebagai landasan fenomena yang terjadi di lapangan khususnya fenomena yang terjadi di daerah Provinsi Riau berdasarkan dari nilai indeks demokrasi, nilai indeks ketenagakerjaan, dan nilai indeks kemiskinan Provinsi Riau.

2. Tinjauan pustaka

2.1. Berita

Berita yang dibahas dalam KBBI berarti sebuah laporan yang cepat tentang hal penting atau peristiwa yang sedang terjadi. Defenisi berita yang dikemukakan oleh para pakar jurnalistik dan komunikasi sebagai berikut [4]:

1. Dean M Lyle Spencer mengatakan berita adalah sebuah ide atau kenyataan yang memiliki kebenaran sehingga pembaca tertarik untuk membacanya.
2. Berita menurut Willard C. Bleyer adalah hal terbaru yang diproses oleh para wartawan untuk dimasukkan ke dalam media agar menarik dan memberikan makna bagi pembacanya.
3. Pendapat William S. Maulsby tentang berita ialah sebuah penuturan yang benar serta tidak memihak, dan memiliki fakta baru yang penting untuk menarik perhatian pembaca surat kabar.
4. Laporan cepat dari suatu kejadian atau peristiwa yang menarik, penting, dan faktual yang menyangkut kepentingan pembacanya disebut berita menurut Michtel V. Charnley.

2.2. Text Mining

Text mining merupakan diantara bidang yang terdapat pada *data mining* [5]. *Text mining* adalah proses memperoleh informasi yang jelas, berguna, dan *eksplisit* dari aktifitas menganalisa terhadap sebuah teks agar dapat dimengerti komputer [6]. Solusi yang diberikan dalam proses *text mining* yaitu menganalisa, pemrosesan, dan pengorganisasian atau pengelompokkan dari data *unstructured* dalam jumlah besar yang bersumber dari sekumpulan dokumen teks yang tidak memiliki format yang terstruktur [2].

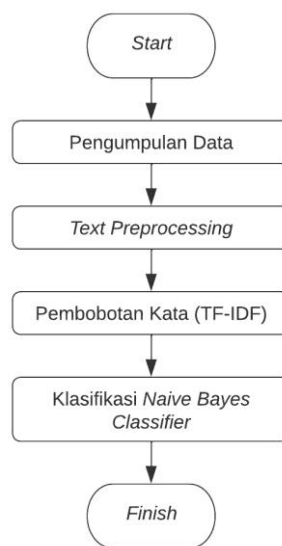
2.3. Penelitian Terkait

Penelitian mengenai topik klasifikasi berita atau teks menggunakan metode *naïve bayes classifier* terus berkembang. Hal ini dibuktikan dengan banyak penelitian mengenai topik tersebut. Berikut beberapa penelitian mengenai proses klasifikasi yang terkait dengan penelitian ini, seperti penelitian yang dilakukan oleh Febry Eka Purwiantono dan Addin Aditya yang berjudul “Klasifikasi Sentimen Sara, Hoaks dan Radikal Pada Postingan Media Sosial Menggunakan Algoritma Naïve Bayes Multinomial Text” berhasil mengklasifikasikan sentiment yang melanggar UU-ITE dengan akurasi 99,62% [7]. Berikutnya penelitian yang berjudul “Klasifikasi Berita Hoax Dengan Menggunakan Metode *Naïve Bayes*” memperoleh hasil akurasi tertinggi sebesar 85.28% pada pengujian 6 *fold cross validation*, sedangkan nilai *precision* 0,896 dan *recall* 0,853 yang dilakukan oleh Hery Mustofa dan Adzhal Arwani Mahfudh [8]. Penelitian selanjutnya memperoleh hasil akurasi sebesar 78.75%, *recall* 80.56%, dan *precision* 78.75% dalam mengklasifikasikan judul berita online menggunakan metode *naïve bayes* yang dilakukan oleh Muhammad Sholih ‘Afif, Moh. Iqbal Alghifari, dan

Muhammad Muzakir dengan judul “Text Mining Untuk Mengklasifikasi Judul Berita Online Studi Kasus Radar Banjarmasin Menggunakan Metode Naïve Bayes” [9]. Berikutnya penelitian dengan judul “Pemanfaatan Metode Klasifikasi *Naïve Bayes* Untuk Pendeteksi Berita *Hoax* Pada Artikel Berbahasa Indonesia” yang dilakukan oleh Soleman memperoleh akurasi 72% dengan penambahan atribut dalam mendeteksi berita hoax pada artikel Bahasa Indonesia [10]. Berdasarkan beberapa contoh penelitian yang terkait dengan penelitian ini, membuktikan bahwa metode *naïve bayes classifier* dapat digunakan dalam proses pengklasifikasian teks ataupun berita serta memiliki nilai akurasi yang cukup baik [11].

3. Metode penelitian

Metode kuantitatif digunakan sebagai landasan dalam melakukan penelitian ini. Penelitian kuantitatif menekankan tentang pengujian teori berdasarkan variable yang diukur dan melakukan analisa terhadap data dengan prosedur statistik [12]. Dalam penelitian kuantitatif terdapat batasan dalam lingkup penelitian yang membatasi kebutuhan variabel dan populasi pada suatu penelitian. Penelitian kuantitatif dilakukan dengan rancangan tahapan yang terstruktur dan sesuai dengan sistematika penelitian ilmiah. Berikut tahapan yang akan dilakukan dalam penelitian ini pada Gambar 1 di bawah ini.



Gambar 1. Tahapan Penelitian

Berikut penjelasan tahapan penelitian sesuai gambar diatas:

a. Pengumpulan Data

Dalam penelitian ini mengumpulkan data secara manual dari tanggal 30 Januari 2022 sampai dengan 10 Maret 2022 dari beberapa portal berita khusus daerah Riau. Data yang dikumpulkan berjumlah 510 data berita, selanjutnya dilakukan proses pelabelan secara manual oleh para ahli dari bidang sosial di Badan Pusat Statistik Provinsi Riau. Dari 510 berita tersebut, sebanyak 170 berita berkategori demokrasi, 170 berita berkategori kemiskinan, dan 170 berita berkategori ketenagakerjaan. Dokumen berita disimpan dalam bentuk format *google spreadsheet*. Berikut ini merupakan contoh *dataset* berita yang telah dikumpulkan dan dilabeli:

Dataset Berita	Kelas
Rapat Paripurna dengan agenda pelantikan Pergantian Antar Waktu (PAW) Wakil Ketua DPRD	Demokrasi
Pemerintah Kota (Pemko) Pekanbaru sedang mengumpulkan data masyarakat miskin positif Covid-19	Kemiskinan
Himbauan kepada Badan Operasi Bersama (BOB) PT Bumi Siak Pusako (BSP)-Pertamina Hulu, agar menunda pengurangan tenaga kerja.	Ketenagakerjaan

Tabel 1. *Dataset* Berita

Dataset yang digunakan dibagi dengan perbandingan yaitu data latih 70% : 30% data uji, data latih 80% : 20% data uji, dan data latih 90% : 10% data uji. Berikut rincian perbandingan *dataset* yang digunakan.

Data Berita	Kelas	Pembagian data latih dan data uji					
		Data Latih	Data Uji	Data Latih	Data Uji	Data Latih	Data Uji
		70%	30%	80%	20%	90%	10%
510	Demokrasi	119	51	136	34	153	17
	Kemiskinan	119	51	136	34	153	17
	Ketenagakerjaan	119	51	136	34	153	17
Jumlah		357	153	408	102	459	51

Tabel 2. Skenario Pembagian Data

b. *Text Preprocessing*

Setelah data siap dikumpulkan, selanjutnya data melalui proses *text preprocessing* yang penting untuk dilakukan dalam proses pengklasifikasian data. Tahapan *text preprocessing* akan membersihkan data berita dari komponen yang tidak diperlukan, membentuk pola kata yang sama dan mengurangi volume dari kata agar mempermudah proses klasifikasi. Agar memperoleh dimensi kata yang lebih kecil sehingga lebih efisien dan presisi dalam proses komputasi [13]. Masalah terbesar dalam proses klasifikasi adalah menjaga dimensi setiap kata [14]. Berikut adalah tahapan yang harus dilakukan dalam proses *text preprocessing*:

1. *Cleaning*

Data yang tidak digunakan akan dibersihkan seperti tanda baca, angka, symbol, URL, dan emoticon.

2. *Case Folding*

Tahapan *case folding* seluruh huruf yang terdapat pada data akan diubah menjadi huruf kecil (*lowercase*).

3. *Tokenizing*

Tahapan *tokenizing* (*tokenisasi*) akan memisahkan kalimat menjadi potongan-potongan kata atau disebut token (potongan kata tunggal).

4. *Normalisasi*

Tahapan normalisasi dilakukan perubahan kata tidak baku atau salah ejaannya menjadi kata yang baku dengan menggunakan kamus normalisasi yang dibuat manual berdasarkan pengecekan data secara manual.

5. *Removal Stopword*

Tahapan *removal stopwords* menghapus seluruh kata yang tidak dibutuhkan dalam proses klasifikasi dari hasil normalisasi sebelumnya. Proses penghapusan kata-kata tersebut berdasarkan filtering menggunakan library NLTK untuk bahasa Indonesia dan kamus *stopword* yang dibuat sendiri secara manual.

6. *Stemming*

Tahapan *stemming* akan menganalisa setiap kata dari hasil *removal stopwords* sebelumnya untuk memperoleh suatu kata dasar berdasarkan kata-kata tersebut. Proses *stemming* akan dihilangkan imbuhan awalan kata, akhiran kata, dan sisipan kata, serta *confixes* (kombinasi awalan dan akhiran) sesuai panduan KBBI. Pada proses *stemming* ini menggunakan kelas *StemmerFactory* dari library Sastrawi.

c. *Pembobotan Kata (TF-IDF)*

Tahapan ini akan memberi bobot pada setiap kata untuk menghitung berapa kali sebuah kata muncul dalam setiap dokumen disebut *Term Frequency (TF)* dan memiliki persamaan sebagai berikut:

$$tf_{ij} = \frac{f_a}{\max f_a(j)}$$

Keterangan

tf_{ij} : Jumlah kemunculan kata (*term*) dalam setiap dokumen

f_a : Frekuensi terdapat kata (*term*) i dalam dokumen j

$\max f_a(j)$: Total kata (*term*) pada dokumen j

Selanjutnya untuk menghitung kata tertentu yang banyak muncul pada sebuah dokumen disebut *Inverse Document Frequency (IDF)* dengan persamaan sebagai berikut:

$$idf_{ij} = \ln\left(\frac{D}{df_i}\right) + 1$$

Keterangan

idf_{ij} : Total munculnya kata (*term*) dalam sebuah dokumen

D : Total semua dokumen

df_i : Total dokumen yang terdapat kata (*term*) didalamnya

Term Frequency Inverse Document Frequency (TF-IDF) menjadi tolak ukur perhitungan statistik dalam proses analisa seberapa penting keberadaan sebuah kata pada sebuah dokumen yang dapat diasumsikan sebagai berikut:

$$w_{ij} = tf_{ij} \times idf_{ij}$$

Keterangan

- w_{ij} : Bobot kata (*term*) terhadap dokumen
- tf_{ij} : Total munculnya kata (*term*) dalam setiap dokumen
- idf_{ij} : Total munculnya kata (*term*) dalam sebuah dokumen

d. Klasifikasi *Naïve Bayes Classifier*

Naïve bayes classifier (NBC) merupakan model yang menggunakan metode statistik dan probabilitas dalam proses klasifikasi yang dikenalkan oleh Thomas Bayes (ilmuwan Inggris), yaitu tentang prediksi probabilitas yang terjadi dimasa depan dengan mempertimbangkan pengalaman dari masa yang telah terjadi [15]. Rumus Teorema Bayes memiliki persamaan sebagai berikut[16]:

$$P(Y|Z) = \frac{P(Y)P(Z|Y)}{P(Z)}$$

Keterangan

- $P(Y)$: Probabilitas kejadian Y
- $P(Z)$: Probabilitas kejadian Z
- $P(Y|Z)$: Probabilitas kejadian Y berdasarkan kejadian Z
- $P(Z|Y)$: Probabilitas kejadian Z berdasarkan kejadian Y

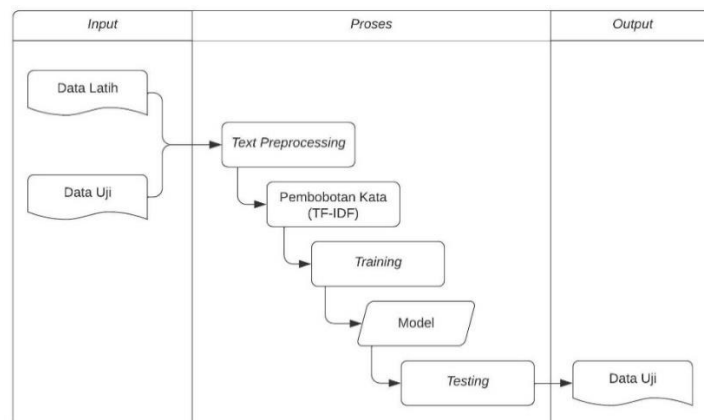
Dalam menghitung probabilitas yang paling tinggi dari suatu kategori pada dokumen yang diujikan dalam proses klasifikasi dapat diasumsikan sebagai berikut:

$$V_{MAP} = \underset{v_j \in V}{argmax} P(v_j) \prod_i P(a_i|v_j)$$

Keterangan

- V_{MAP} : Probabilitas palingtinggi dari seluruh kategori
- $P(v_j)$: Probabilitas dari sekumpulan data apakah terdapat kategori tertentu
- $P(a_i|v_j)$: Probabilitas kemunculan kata a_i dalam sebuah dokumen dengan kategori v_j

Berikut tahapan proses klasifikasi menggunakan metode *naïve bayes classifier* pada Gambar 2 dibawah ini.



Gambar 2. Tahapan Klasifikasi *Naïve Bayes Classifier*

Berikut penjelasan dari Gambar 2 diatas:

1. *Input*

Pada tahapan *input* data menggunakan data berita yang telah dikumpulkan dan dilabeli secara manual oleh tim ahli dari bidang sosial Badan Pusat Statistik Provinsi Riau sebanyak 510 berita. Selanjutnya data berita

akan dibagi menjadi data latih dan data uji sesuai skenario 70% : 30%, 80% : 20%, dan 90% : 10% untuk mendapatkan hasil akurasi terbaik.

2. Proses klasifikasi naïve bayes classifier

Pada proses klasifikasi data diproses melalui tahapan text preprocessing, pembobotan kata (tf-idf), dan pemodelan naïve bayes classifier.

3. Output

Tahap terakhir yaitu hasil data yang telah diklasifikasikan sesuai kelas demokrasi, kemiskinan, dan ketenagakerjaan.

4. Hasil dan pembahasan

Berikut merupakan hasil dari setiap proses yang dilakukan dalam penelitian ini, yaitu proses klasifikasi berita menggunakan metode naïve bayes classifier:

a. Text Preprocessing

1. Cleaning

Berikut hasil tahapan *cleaning* pada Tabel 3 di bawah ini.

Berita sebelum tahapan <i>cleaning</i>	Berita setelah tahapan <i>cleaning</i>
Rapat Paripurnna dengan agenda pelantikan Pergantian Antar Waktu (PAW) Wakil Ketua DPRD	Rapat Paripurnna dengan agenda pelantikan Pergantian Antar Waktu PAW Wakil Ketua DPRD
Pemerintah Kota (Pemko) Pekanbaru sedang mengumpulkan data masyarakat miskin positif Covid-19	Pemerintah Kota Pemko Pekanbaru sedang mengumpulkan data masyarakat miskin positif Covid
Himbauan kepada Badan Operasi Bersama (BOB) PT Bumi Siak Pusako (BSP)-Pertamina Hulu, agar menunda pengurangan tenaga kerja.	Himbauan kepada Badan Operasi Bersama BOB PT Bumi Siak Pusako BSP Pertamina Hulu agar menunda pengurangan tenaga kerja

Tabel 3. Hasil *Cleaning*

2. Case Folding

Berikut hasil tahapan *case folding* pada Tabel 4 di bawah ini.

Berita sebelum tahapan <i>case folding</i>	Berita setelah tahapan <i>case folding</i>
Rapat Paripurnna dengan agenda pelantikan Pergantian Antar Waktu PAW Wakil Ketua DPRD	rapat paripurnna dengan agenda pelantikan pergantian antar waktu paw wakil ketua dprd
Pemerintah Kota Pemko Pekanbaru sedang mengumpulkan data masyarakat miskin positif Covid	pemerintah kota pemko pekanbaru sedang mengumpulkan data masyarakat miskin positif covid
Himbauan kepada Badan Operasi Bersama BOB PT Bumi Siak Pusako BSP Pertamina Hulu agar menunda pengurangan tenaga kerja	himbauan kepada badan operasi bersama bob pt bumi siak pusako bsp pertamina hulu agar menunda pengurangan tenaga kerja

Tabel 4. Hasil *Case Folding*

3. Tokenizing

Berikut hasil *tokenizing* pada Tabel 5 di bawah ini.

D1	D2	D3
rapat paripurnna dengan agenda pelantikan pergantian antar waktu paw wakil ketua dprd	pemerintah kota pemko pekanbaru sedang mengumpulkan data masyarakat miskin positif covid	himbauan kepada badan operasi bersama bob pt bumi siak pusako bsp pertamina

		hulu agar menunda pengurangan tenaga kerja
--	--	---

Tabel 5. Hasil *Tokenizing*

4. *Normalisasi*

Berikut hasil *normalisasi* pada Tabel 6 di bawah ini

D1	D2	D3
rapat paripurna dengan agenda pelantikan pergantian antar waktu paw wakil ketua dprd	pemerintah kota pemko pekanbaru sedang mengumpulkan data masyarakat miskin positif covid	himbauan kepada badan operasi bersama bob pt bumi siak pusako bsp pertamina hulu agar menunda pengurangan tenaga kerja

Tabel 6. Hasil *Normalisasi*

5. *Removal Stopword*

Berikut hasil *removal stopwords* pada Tabel 7 di bawah ini.

D1	D2	D3
rapat paripurna agenda pelantikan pergantian wakil ketua dprd	pemerintah kota pemko pekanbaru mengumpulkan data masyarakat miskin positif covid	himbauan badan operasi bumi siak pusako bsp pertamina hulu menunda pengurangan tenaga kerja

Tabel 7. Hasil *Removal Stopword*

6. *Stemming*

Berikut hasil *stemming* pada Tabel 8 di bawah ini.

D1	D2	D3
rapat paripurna agenda lantik	perintah kota pemko pekanbaru	himbauan badan operasi bumi

ganti wakil ketua dprd	kumpul data masyarakat miskin positif covid	siak pusako bsp pertamina hulu tunda kurang tenaga kerja
---------------------------------	--	--

Tabel 8. Hasil *Stemming*

b. Pembobotan Kata (TF-IDF)

Berikut hasil perhitungan TF-IDF pada Tabel 9 dibawah ini.

Term	TF			DF	IDF	TF-IDF		
	D1	D2	D3			D1	D2	D3
agenda	1	0	0	1	$\ln(4/2)+1 = 1.693$	1.693	0	0
badan	0	0	1	1	$\ln(4/2)+1 = 1.693$	0	0	1.693
bsp	0	0	1	1	$\ln(4/2)+1 = 1.693$	0	0	1.693
bumi	0	0	1	1	$\ln(4/2)+1 = 1.693$	0	0	1.693
covid	0	1	0	1	$\ln(4/2)+1 = 1.693$	0	1.693	0
data	0	1	0	1	$\ln(4/2)+1 = 1.693$	0	1.693	0
dprd	1	0	0	1	$\ln(4/2)+1 = 1.693$	1.693	0	0
ganti	1	0	0	1	$\ln(4/2)+1 = 1.693$	1.693	0	0
himbauan	0	0	1	1	$\ln(4/2)+1 = 1.693$	0	0	1.693
hulu	0	0	1	1	$\ln(4/2)+1 = 1.693$	0	0	1.693
kerja	0	0	1	1	$\ln(4/2)+1 = 1.693$	0	0	1.693
ketua	1	0	0	1	$\ln(4/2)+1 = 1.693$	1.693	0	0
kota	0	1	0	1	$\ln(4/2)+1 = 1.693$	0	1.693	0
kumpul	0	1	0	1	$\ln(4/2)+1 = 1.693$	0	1.693	0
kurang	0	0	1	1	$\ln(4/2)+1 = 1.693$	0	0	1.693
lantik	1	0	0	1	$\ln(4/2)+1 = 1.693$	1.693	0	0
masyarakat	0	1	0	1	$\ln(4/2)+1 = 1.693$	0	1.693	0
miskin	0	1	0	1	$\ln(4/2)+1 = 1.693$	0	1.693	0
operasi	0	0	1	1	$\ln(4/2)+1 = 1.693$	0	0	1.693
paripurnna	1	0	0	1	$\ln(4/2)+1 = 1.693$	1.693	0	0
pekanbaru	0	1	0	1	$\ln(4/2)+1 = 1.693$	0	1.693	0
pemko	0	1	0	1	$\ln(4/2)+1 = 1.693$	0	1.693	0
perintah	0	1	0	1	$\ln(4/2)+1 = 1.693$	0	1.693	0
pertamina	0	0	1	1	$\ln(4/2)+1 = 1.693$	0	0	1.693
positif	0	1	0	1	$\ln(4/2)+1 = 1.693$	0	1.693	0
pusako	0	0	1	1	$\ln(4/2)+1 = 1.693$	0	0	1.693
rapat	1	0	0	1	$\ln(4/2)+1 = 1.693$	1.693	0	0
siak	0	0	1	1	$\ln(4/2)+1 = 1.693$	0	0	1.693
tenaga	0	0	1	1	$\ln(4/2)+1 = 1.693$	0	0	1.693
tunda	0	0	1	1	$\ln(4/2)+1 = 1.693$	0	0	1.693
wakil	1	0	0	1	$\ln(4/2)+1 = 1.693$	1.693	0	0

Tabel 9. Hasil TF-IDF

c. Klasifikasi *Naïve Bayes Classifier*

Data berita yang telah diproses pada tahapan text preprocessing dan masuk ke tahapan pembobotan kata (tf-idf), selanjutnya akan melalui proses klasifikasi menggunakan metode naïve bayes classifier untuk menghasilkan model yang dapat mengklasifikasikan data berita baru tanpa pelabelan manual. Dalam penelitian ini terdapat 3 skenario dalam pembagian data. Berikut hasil klasifikasi berita dengan skenario pembagian data dengan data latih 70% dan 30% data uji pada Gambar 3 dibawah ini.

```

Accuracy of the classifier is : 0.8954248366013072

Confusion matrix:
[[36  3  4]
 [ 1 50  3]
 [ 3  2 51]]

Classification report is:
              precision    recall  f1-score   support

     0       0.90      0.84      0.87         43
     1       0.91      0.93      0.92         54
     2       0.88      0.91      0.89         56

 accuracy          0.90
 macro avg          0.90
 weighted avg       0.90
    
```

Gambar 3. Hasil Klasifikasi Berita Dengan Data Latih 70% dan 30% Data Uji

Selanjutnya untuk hasil klasifikasi berita dengan skenario pembagian data dengan data latih 80% dan 20% data uji pada Gambar 4 dibawah ini.

```

Accuracy of the classifier is : 0.9019607843137255

Confusion matrix:
[[25  1  1]
 [ 2 29  1]
 [ 2  3 38]]

Classification report is:
              precision    recall  f1-score   support

     0       0.86      0.93      0.89         27
     1       0.88      0.91      0.89         32
     2       0.95      0.88      0.92         43

 accuracy          0.90
 macro avg          0.90
 weighted avg       0.90
    
```

Gambar 4. Hasil Klasifikasi Berita Dengan Data Latih 80% dan 20% Data Uji

Dan untuk hasil klasifikasi berita dengan skenario pembagian data dengan data latih 90% dan 10% data uji pada Gambar 5 dibawah ini.

```

Accuracy of the classifier is : 0.9411764705882353

Confusion matrix:
[[21  0  1]
 [ 0 16  1]
 [ 1  0 11]]

Classification report is:
              precision    recall  f1-score   support

     0       0.95      0.95      0.95         22
     1       1.00      0.94      0.97         17
     2       0.85      0.92      0.88         12

 accuracy          0.94
 macro avg          0.93
 weighted avg       0.94
    
```

Gambar 5. Hasil Klasifikasi Berita Dengan Data Latih 90% dan 10% Data Uji

d. Pengujian *Confusion Matrix*

Pengujian yang dilakukan dalam penelitian ini berdasarkan data yang telah dibagi menjadi data latih dan data uji dari data berita yang berjumlah 510 berita. Pembagian data latih dan data uji dalam penelitian ini

menggunakan perbandingan data latih 70% : 30% data uji, data latih 80% : 20% data uji, dan data latih 90% : 10% data uji.

1. Pengujian terhadap data latih 70% dan 30% data uji

Berikut hasil pengujian *confusion matrix* untuk data latih 70% dan 30% data uji pada Tabel 10 di bawah ini.

<i>Predicted Classification</i>	<i>Actual Classification</i>		
	Demokrasi Positive	Kemiskinan Positive	Ketenagakerjaan Positive
Demokrasi Negative	36	3	4
Kemiskinan Negative	1	50	3
Ketenagakerjaan Negative	3	2	51

Tabel 10. Hasil *Confusion Matrix* pada data latih 70% dan 30% data uji

$$\begin{aligned} \text{Perhitungan akurasi} &= \frac{36+50+51}{36+3+4+1+50+3+3+2+51} \times 100\% \\ &= \frac{137}{153} \times 100\% \\ &= 90\% \end{aligned}$$

2. Pengujian terhadap data latih 80% dan 20% data uji

Berikut hasil pengujian *confusion matrix* untuk data latih 80% dan 20% data uji pada Tabel 11 di bawah ini.

<i>Predicted Classification</i>	<i>Actual Classification</i>		
	Demokrasi Positive	Kemiskinan Positive	Ketenagakerjaan Positive
Demokrasi Negative	25	1	1
Kemiskinan Negative	2	29	1
Ketenagakerjaan Negative	2	3	38

Tabel 11. Hasil *Confusion Matrix* pada data latih 80% dan 20% data uji

$$\begin{aligned} \text{Perhitungan akurasi} &= \frac{25+29+38}{25+1+1+2+29+1+2+3+38} \times 100\% \\ &= \frac{92}{102} \times 100\% \\ &= 90\% \end{aligned}$$

3. Pengujian terhadap data latih 90% dan 10% data uji

Berikut hasil pengujian *confusion matrix* untuk data latih 90% dan 10% data uji pada Tabel 12 di bawah ini.

<i>Predicted Classification</i>	<i>Actual Classification</i>		
	Demokrasi Positive	Kemiskinan Positive	Ketenagakerjaan Positive
Demokrasi Negative	21	0	1
Kemiskinan Negative	0	16	1
Ketenagakerjaan Negative	1	0	11

Tabel 12. Hasil *Confusion Matrix* pada data latih 90% dan 10% data uji

$$\begin{aligned} \text{Perhitungan akurasi} &= \frac{21+16+11}{21+0+1+0+16+1+1+0+11} \times 100\% \\ &= \frac{48}{51} \times 100\% \\ &= 94\% \end{aligned}$$

5. Kesimpulan

Berdasarkan hasil yang diperoleh dalam penelitian ini, dapat disimpulkan bahwa:

1. Metode *naïve bayes classifier* terbukti dapat digunakan dalam proses klasifikasi berita.
2. Proses klasifikasi menggunakan metode *naïve bayes classifier* menghasilkan akurasi tertinggi yaitu 94% dengan pembagian data latih 90% dan 10% data uji dari *dataset* berita yang digunakan.

Daftar Pustaka

- [1] A. Nurhadi, "Implementasi Algoritma Naïve Bayes Classifier Berbasis Particle Swarm Optimization (PSO) Untuk Klasifikasi Konten Berita Digital Bahasa Indonesia," *J. Speed – Sentra Penelit. Eng. dan Edukasi*, vol. 8, no. 3, pp. 48–56, 2016.
- [2] F. Nurhuda, S. Widya Sihwi, and A. Doewes, "Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier," *J. Teknol. Inf. ITSmart*, vol. 2, no. 2, p. 35, 2016.
- [3] T. Arifin and D. Ariesta, "Prediksi Penyakit Ginjal Kronis Menggunakan Algoritma Naive Bayes Classifier Berbasis Particle Swarm Optimization," *J. Tekno Insestif*, vol. 13, no. 1, pp. 26–30, 2019.
- [4] M. S. Restendy, "DAYA TARIK JURNALISTIK, PERS, BERITA DAN PERBEDAAN PERAN DALAM NEWS CASTING," *al-Hikmah*, vol. 4, no. 2, pp. 31–48, 2016.
- [5] D. Ariyanti and K. Iswardani, "Teks Mining untuk Klasifikasi Keluhan Masyarakat Pada Pemkot Probolinggo Menggunakan Algoritma Naïve Bayes," *J. IKRA-ITH Inform.*, vol. 4, no. 3, pp. 125–132, 2020.
- [6] A. Taufik, "Optimasi Particle Swarm Optimization Sebagai Seleksi Fitur Pada Analisis Sentimen Review Hotel Berbahasa Indonesia Menggunakan Algoritma Naïve Bayes," *J. Tek. Komput.*, vol. III, no. 2, pp. 40–47, 2017.
- [7] F. E. Purwiantono and A. Aditya, "Klasifikasi Sentimen Sara, Hoaks Dan Radikal Pada Postingan Media Sosial Menggunakan Algoritma Naive Bayes Multinomial Text," *J. Tekno Kompak*, vol. 14, no. 2, p. 68, 2020.
- [8] H. Mustofa and A. A. Mahfudh, "Klasifikasi Berita Hoax Dengan Menggunakan Metode Naive Bayes," *Walisongo J. Inf. Technol.*, vol. 1, no. 1, p. 1, 2019.
- [9] M. Sholih 'afif, M. Muzakir, M. I. Al, and G. Al Awalaien, "Text Mining Untuk Mengklasifikasi Judul Berita Online Studi Kasus Radar Banjarmasin Menggunakan Metode Naïve Bayes," *Kumpul. J. Ilmu Komput.*, vol. 08, no. 2, pp. 199–208, 2021.
- [10] S. Soleman, "Pemanfaatan Metode Klasifikasi Naïve Bayes Untuk Pendeteksi Berita Hoax Pada Artikel Berbahasa Indonesia," *J. CoreIT J. Has. Penelit. Ilmu ...*, vol. 7, no. 2, pp. 82–93, 2021.
- [11] E. T. Handayani and A. Sulistiyawati, "Analisis Sentimen Respon Masyarakat Terhadap Kabar Harian Covid-19 Pada Twitter Kementerian Kesehatan," *J. Teknol. dan Sist. Inf.*, vol. 2, no. 3, pp. 32–37, 2021.
- [12] C. R. Semiawan, *Metode Penelitian Kuantitatif*. 2017.
- [13] B. S. Prakoso, D. Rosiyadi, H. S. Utama, and D. Aridarma, "Klasifikasi Berita Menggunakan Algoritma Naive Bayes Classifier Dengan Seleksi Fitur Dan Boosting," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 2, pp. 227–232, 2019.
- [14] Apif Supriadi and Fatmasari, "Implementasi Metode Klasifikasi Naive Bayes Pada Sistem Analisis Opini Pengguna Twitter Berbasis Web," *J. Sist. Inf.*, vol. 10, no. 1, pp. 46–54, 2021.
- [15] I. Rasila and U. Ristian, "Implementasi Metode Naive Bayes Classifier Pada Sistem Pengklasifikasi Berita Otomatis Berbasis Website (Studi Kasus: Berita Lokal Dari Mediamassa Online Kalimantan Barat)," *Coding J. Komput. dan Apl.*, vol. 07, no. 2, pp. 49–60, 2019.
- [16] H. A. R. Harpizon, R. Kurniawan, I. Iskandar, R. Salambue, E. Budianita, and F. Syafria, "Analisis Sentimen Komentar Di YouTube Tentang Ceramah Ustadz Abdul Somad Menggunakan Algoritma Naïve Bayes," *Anal. Sentimen Komentar Di YouTube Tentang Ceramah Ustadz Abdul Somad Menggunakan Algoritm. Naïve Bayes*, vol. 5, no. 1, pp. 131–140, 2022.