



Jurnal Teknologi Informasi dan Komunikasi

Vol: 14 No 01 2023

E-ISSN: 2477-3255

Diterima Redaksi: 25-03-2023 | Revisi: 01-04-2023 | Diterbitkan: 28-05-2023

Implementation Of Text Mining And Pattern Discovery With Naive Bayes Algorithm For Classification Of Text Documents

Novia Lestari¹, Ozzy Secio Riza², Reno Ardinal³

^{1,2,3}Program Studi Sistem Informasi Universitas Islam Negeri Imam Bonjol Padang

^{1,2,3}Balai Gadang, Kecamatan Koto Tangah, Kota Padang, Sumatera Barat

e-mail: ¹noviia.lestari@uinib.ac.id, ²ozzysecioriza@uinib.ac.id, ³renoardinal@gmail.com

Abstract

Classification of text documents can be managed manually by using human-made classification rules. However, as many text document files exist today, the application of machine learning can help to classify the documents more effectively and the structured. Data mining with the Naïve Bayes algorithm can help the process of searching for a set of patterns or characteristics that explain and separate a classification of data based on the aim that the model can used to predict and classify the the data that has been used. This study uses text mining and pattern discovery techniques with the naïve Bayes algorithm used in the Indonesian language online news classification process with an accuracy test result of 63.9 and a low error rate of 41.02%.

Keywords: Text Mining, Pattern Discovery, Naïve Bayes Algorithm, Classification

Implementasi Text Mining Dan Pattern Discovery Dengan Algoritma Naive Bayes Untuk Klasifikasi Dokumen Teks

Abstrak

Klasifikasi dokumen teks dapat dilakukan secara manual menggunakan aturan klasifikasi yang dibuat manusia. Namun, seiring banyaknya file dokumen teks yang ada saat ini, penerapan machine learning dapat membantu pengklasifikasian dokumen secara lebih efektif dan terstruktur. Data mining dengan algoritma Naive Bayes dapat membantu menemukan sekumpulan pola atau fitur yang menggambarkan dan memisahkan kelas data, sehingga pola tersebut dapat digunakan untuk memprediksi dan mengklasifikasikan kelas dari suatu objek yang tidak diketahui kelasnya berdasarkan informasi dari kelas yang ada sebelumnya. Penelitian ini menggunakan teknik text mining dan pattern discovery dengan algoritma naïve bayes yang digunakan dalam proses klasifikasi berita online berbahasa Indonesia dengan hasil pengujian nilai akurasi sebesar 63,9 dan error rate yang rendah yaitu sebesar 41,02%.

Kata kunci: Text Mining, Pattern Discovery, Algoritma Naïve Baye, Klasifikasi

<https://doi.org/10.31849/digitalzone.v14i1.13596>

Digital Zone is licensed under a Creative Commons Attribution International (CC BY-SA 4.0)

1. Pendahuluan

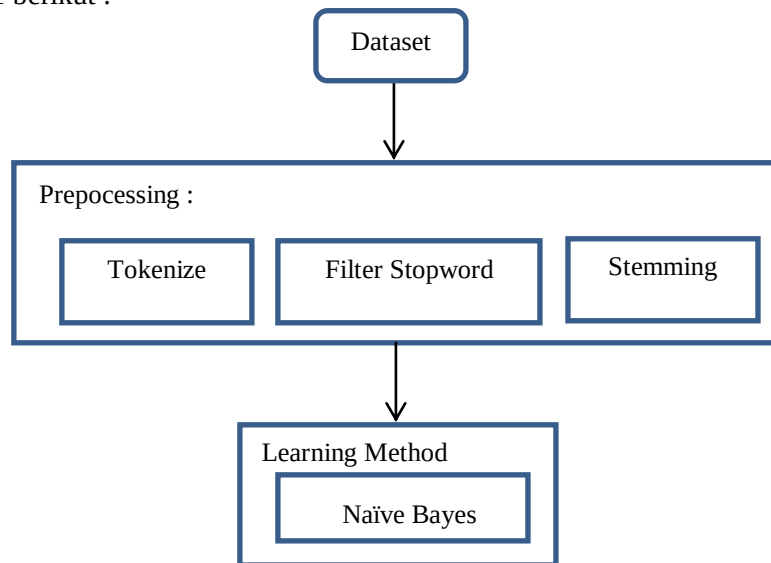
Peningkatan aliran informasi di era big data seperti saat sekarang ini menyebabkan terjadinya penumpukan data yang tidak dapat dihindari. Data yang berlimpah yang dihasilkan dari peningkatan aliran informasi tersebut membutuhkan teknik dan metode yang tepat agar dapat dimanfaatkan lagi sebagai pengetahuan baru yang berguna bagi pemakainya [1]. Ketepatan dalam pemilihan metode, teknik, atau algoritma mempengaruhi tujuan dan proses penambangan data dari sebuah dokumen [2]. Penggalian informasi dan pengetahuan baru dari kumpulan data yang jumlahnya besar ini dapat dilakukan dengan mengekstrak dan mencari pola-pola data menggunakan teknik Data mining [3]. Data mining dapat mengekstrak informasi dari data-data yang kompleks dan sulit dianalisis secara mendalam sekalipun, seperti kebiasaan/perilaku belajar siswa yang sering dijadikan objek penelitian dalam *Educational Data Mining* [4]. Dalam dunia bisnis dan perdagangan, data mining menjadi teknik yang handal yang dapat membantu perusahaan dalam menggali informasi penting tentang perilaku pelanggan, sehingga bisa menentukan pelanggan potensial yang bisa memberi keuntungan lebih bagi perusahaan [5]. Klasifikasi merupakan metode data mining yang populer yang bisa diterapkan untuk berbagai subjek maupun objek data, salah satunya dalam mengklasifikasikan dokumen teks. Klasifikasi dokumen teks dapat dilakukan secara manual menggunakan aturan klasifikasi yang dibuat manusia. Namun, seiring banyaknya file dokumen teks yang ada saat ini, penerapan *machine learning* dapat membantu pengklasifikasian dokumen secara lebih efektif dan terstruktur [6], terlebih untuk dokumen teks berita online yang digunakan sebagai sumber data dalam penelitian ini. Mencari dan mengklasifikasikan teks secara manual membutuhkan waktu dan tenaga ketika banyak dokumen yang harus diklasifikasi. Oleh karena itu, diperlukan suatu metode untuk membantu mengklasifikasikan dokumen secara cepat dan akurat. Kebutuhan ini dipermudah dengan adanya klasifikasi dalam sebuah dokumen teks. Klasifikasi adalah proses menemukan sekumpulan pola atau fungsi yang mendeskripsikan dan membedakan kelas dari suatu data sehingga pola yang ditemukan tersebut tersebut dapat digunakan untuk memprediksi kelas dari objek yang tidak diketahui. Klasifikasi melibatkan dua proses, yaitu membangun model klasifikasi dari kelas data yang telah ditentukan (dataset pelatihan) dan menggunakan model ini untuk mengklasifikasikan data tekstual dan mengukur akurasi model. [7]. Klasifikasi dokumen secara otomatis dengan cara mengelompokkan suatu dokumen ke dalam suatu kategori berdasarkan analisis data training atau data testing [8].

Algoritma Naive Bayes adalah metode klasifikasi probabilistik sederhana yang digunakan untuk menghitung atau mencari probabilitas maksimum dari pengklasifikasian data uji ke dalam kelas yang sesuai. Pada penelitian sebelumnya, dilakukan klasifikasi konten berita menggunakan algoritma *Naive Bayes* dengan nilai akurasi yaitu 73,2% [9]. Analisis sentiment dari komentar yang bersumber dari sebuah website juga dapat diklasifikasikan dengan *Naive bayes classifier* dengan tepat [10].

Berdasarkan rujukan penelitian yang ada tentang data mining metode klasifikasi dengan *Naive Bayes Classifier* di atas, dengan menerapkan metode *text mining* dan *pattern discovery* pada penelitian ini, dilakukan pengklasifikasian berita online berbahasa Indonesia sesuai dengan kategori yang ada yaitu edukasi, bisnis dan keuangan, dan olahraga berdasarkan tahapan pembelajaran dari dokumen latih (*data training*) yang ada. Setelah tahapan pembelajaran, dilakukan uji terhadap dokumen testing untuk menentukan klasifikasi dari berita berbahasa Indonesia tersebut, dimana dari hasil uji ini dapat dilihat juga nilai akurasi dari algoritma *naive bayes* dalam mengklasifikasikan berita online yang terdiri dari 300 data berupa artikel berita online berbahasa Indonesia yang dibagi menjadi data training dan data testing untuk mengelompokkan berita sesuai dengan kategori yang ada.

2. Metode Penelitian

Penelitian ini dilakukan dalam 3 tahap, yaitu pengumpulan dataset, *text mining*, dan *pattern discovery* (penemuan pola) dengan algoritma *Naïve Bayes Classifier* dapat dilihat pada gambar 1 berikut :



Gambar 1. Metode Penelitian

Tahapan-tahapan yang dilakukan berdasarkan gambar metode penelitian di atas adalah sebagai berikut :

1. Pengumpulan Dataset

Dataset merupakan data yang digunakan dalam proses *text mining* dan kemudian dilanjutkan dengan *pattern discovery* (penemuan pola) dengan algoritma *Naïve Bayes*. Dalam penelitian ini, kumpulan berita online Indonesia digunakan sebagai dokumen pelatihan (*training data*) dan dokumen pengujian (*test data*) dari situs resmi Media Indonesia yaitu www.mediaindonesia.com yang sering diakses oleh pencari berita baik di dalam negeri maupun luar negeri. Data berita yang digunakan terdiri dari 3 kategori yaitu bisnis dan keuangan, edukasi, dan olahraga. Dengan 200 data sebagai data latih dan 100 data sebagai data uji yang diambil secara acak.

2. Tahapan *Preprocessing*

Proses yang dilakukan pada tahapan *processing* seperti *tokenizing* yaitu menghilangkan simbol dan tanda baca yang ada pada dokumen berita. Setelah itu dilakukan proses *parsing* atau *filter stopwords* yaitu teks dibagi menjadi sekumpulan kata tanpa memperhatikan hubungan antara kata dan peran atau posisinya dalam kalimat dengan cara menghilangkan kata penghubung seperti “yang”, “dan”, “di”, “antara”, dll. Kemudian dilakukan proses *stemming* yaitu pengecekan atau pencarian kata sesuai kata dasarnya, salah satu caranya dengan menghilangkan awalan dan imbuhan yang terdapat pada kata [11].

3. *Learning Method*

Algoritma *Naïve Bayes Classifier* digunakan pada tahap *pattern discovery* (penemuan pola). Ada 2 proses yang dilakukan pada tahapan ini, yaitu *learn naïve bayes* untuk membentuk informasi dalam bentuk probabilitas dan *classify naïve bayes* untuk mengembalikan perkiraan nilai target dari dokumen yang diklasifikasikan. Pada fase pembelajaran Naive Bayes, proses berlangsung sehingga pengetahuan kata yang ada diperiksa untuk setiap kata yang muncul dalam data pelatihan. Jika kata ditemukan, tambahkan jumlah kemunculan kata sesuai kategori yang ada, jika tidak, tambahkan kata baru dan jumlah kemunculan kata sesuai kategori yang ada. Setelah menambahkan semua kata dan frekuensinya ke pengetahuan kata, hitung probabilitasnya dengan rumus 1 [12] :

$$P(V_j) = \frac{|docs_j|}{|contoh|} \quad (1)$$

dimana $P(V_j)$ adalah probabilitas kemunculan $|docs_j|$ merupakan jumlah dokumen yang ada pada kategori j dan $|Contoh|$ merupakan jumlah dokumen yang digunakan dalam tahapan pelatihan. Setelah itu, cari perkalian probabilitas kata dilihat dari himpunan kategori dengan rumus 2 sebagai berikut :

$$P(W_k|V_j) = \frac{n_k+1}{n+|kosa\ kata|} \quad (2)$$

Dimana P adalah probabilitas kata (W_k) yang dilihat dari V_j (himpunan kategori berita, seperti olahraga, sains, edukasi dan sebagainya). Sedangkan n_k adalah jumlah kemunculan kata W_k pada kategori V_j , n merupakan jumlah semua kata pada kategori V_j dan $|kosakata|$ yaitu jumlah kata yang unik (*distinc*) pada *data training* [13].

Pada tahapan *Classify naïve bayes* dicari nilai probabilitas tertinggi untuk menentukan kelompok data uji ke dalam kelas yang paling sesuai. Langkahnya yaitu : kata yang ada pada data uji, dicari ke dalam pengetahuan kata. Jika kata tersebut ditemukan, cari nilai probabilitasnya dengan rumus (2), jika tidak ada, maka abaikan kata tersebut. Untuk setiap kategori, hitung perkalian probabilitas (1) dan (2). Setelah mendapatkan hasil perkalian probabilitas, bandingkan hasilnya dari semua kategori yang ada, nilai tertinggi yang didapat, maka data dokumen yang menjadi data testing masuk ke kategori dengan nilai tertinggi tersebut [14].

Untuk menguji keakuratan hasil klasifikasi, digunakan Confusion Matrix yang bekerja dengan cara membandingkan hasil klasifikasi dengan penerapan algoritma yang dilakukan dengan hasil klasifikasi yang seharusnya sehingga menghasilkan informasi tingkat akurasi penerapan algoritma tersebut. Tabel confusion matrix dapat dilihat pada Tabel 1 berikut :

Tabel 1. Confusion Matrix		
Klasifikasi Benar	Klasifikasi	
	Prediksi Benar	Prediksi Salah
Aktual Positif	True Positive	False Negative
Aktual Negatif	False Positive	True Negative

Berdasarkan Tabel 1 di atas, *True Positive* (TP) menunjukkan nilai positif dari jumlah *record* data positif yang dilakukan klasifikasi. *False Positives* (FP) menunjukkan nilai positif dari jumlah data *record* negatif yang dilakukan klasifikasi. *False Negative* (FN) menunjukkan nilai negative dari jumlah *record* data positif yang diklasifikasi, sedangkan *True Negative* (TN) menunjukkan nilai negative berdasarkan jumlah *record* data negatif yang diklasifikasi. Untuk memvalidasi kinerja algoritma klasifikasi dapat digunakan metode *K-Fold Cross Validation* yang memisahkan dataset secara acak sebanyak K bagian, kemudian proses klasifikasi akan dilakukan sebanyak k kali percobaan dan hasil evaluasi pun memiliki jumlah nilai K yang sama, dimana masing-masing percobaan menggunakan data partisi ke- K sebagai data proses pembelajaran (*training*) dan sisanya sebagai data validasi (*testing*).

3. Hasil Dan Pembahasan

Misalkan ada 3 kategori yaitu kategori edukasi, olahraga, dan kesehatan. Dokumen yang digunakan pada proses klasifikasi terdiri dari 2 buah dokumen, dimana dokumen pertama sampai dokumen dua merupakan dokumen yang akan digunakan pada tahap pembelajaran yang disebut dengan *data training*, sedangkan dokumen ketiga akan digunakan sebagai dokumen *testing* yang akan diklasifikasikan.

Dokumen 1 (yang sudah di preprocessing) :

(Edukasi)

Judul :[Kapan Ya, Biaya Kuliah Sesuai Kemampuan ?](#)

bandung kompas com perguruan negeri ptn indonesia idealnya menetapkan standar pembiayaan studi mahasiswa kemampuan orangtua bertujuan mendukung aspek keadilan meringankan beban ekonomi mahasiswa berprestasi keluarga idealnya perguruan

universitas indonesia universitas brawijaya menerapkan kewajiban universitas padjadjaran sebatas imbauan rektor universitas padjadjaran ganjar kurnia bandung jawa kemarin ganjar standar pembedaan biaya efektif aspek keadilan kemudahan studi mahasiswa

Kategori = Edukasi

Daftar kata contoh dari dokumen 1 tertera pada Tabel 2 berikut :

Tabel 2. Daftar Kata Contoh Dokumen Pembelajaran 1

No.	Kata	Frekuensi	No.	Kata	Frekuensi
1	kompas	1	20	Beban	1
2	com	1	21	Ekonomi	1
3	perguruan	2	22	Prestasi	1
4	negeri	1	23	Keluarga	1
5	ptn	1	24	Universitas	4
6	indonesia	2	25	Brawijaya	1
7	ideal	2	26	Wajib	1
8	tetap	2	27	padjadjaran	2
9	standar	2	28	Sebatas	1
10	biaya	2	29	Imbau	1
11	studi	2	30	Rector	1
12	mahasiswa	3	31	Ganjar	2
13	mampu	1	32	Kurnia	1
14	orang	1	33	Bandung	2
15	tua	1	34	Jawa	1
16	dukung	1	35	Kemarin	1
17	aspek	2	36	Beda	1
18	adil	2	37	Efektif	1
19	ringan	1	38	Mudah	1
		Jumlah			55

Perhitungan $P(V_j)$ (1)

Diketahui : Jumlah dokumen edukasi = 1
 Jumlah dokumen bisnis keuangan = 0
 $P(\text{edukasi}) = 1/1 = 1$
 $P(\text{bisnis dan keuangan}) = 0/1 = 0$

Perhitungan $P(W_k | V_j)$ (2)

Diketahui : Frekuensi edukasi = 55
 Frekuensi bisnis dan keuangan = 0
 Jumlah kata = 38

Pada kategori edukasi :

$P(\text{kompas} | \text{edukasi}) = (1+1) / (55+38) = 0.022$
 $P(\text{perguruan} | \text{edukasi}) = (2+1) / (55+38) = 0.032$
 $P(\text{ptn} | \text{edukasi}) = (1+1) / (55+38) = 0.022$
 Hasil perhitungan lengkap dapat dilihat dalam Tabel 2

Pada kategori bisnis & keuangan :

$P(\text{kompas} | \text{bisnis keuangan}) = (0+1) / (0+38) = 0.026$
 $P(\text{perguruan} | \text{bisnis keuangan}) = (0+1) / (0+38) = 0.026$
 $P(\text{ptn} | \text{bisnis \& keuangan}) = (0+1) / (0+38) = 0.026$
 Hasil perhitungan lengkap dapat dilihat dalam Tabel 2

Maka pengetahuan kata dokumen 1 tertera pada Tabel 3 berikut :

<https://doi.org/10.31849/digitalzone.v14i1.13596>

Digital Zone is licensed under a Creative Commons Attribution International (CC BY-SA 4.0)

Tabel 3. Daftar Pengetahuan Kata Contoh Dokumen Pembelajaran 1

No	Kata	Frek Edukasi	Frek Bisnis Keuangan	Probabilitas Edukasi	Probabilitas Bisnis Keuangan
1	kompas	1	0	0,022	0,026
2	com	1	0	0,022	0,026
3	perguruan	2	0	0,032	0,026
4	negeri	1	0	0,022	0,026
5	ptn	1	0	0,022	0,026
6	indonesia	2	0	0,032	0,026
7	ideal	2	0	0,032	0,026
8	tetap	2	0	0,032	0,026
9	standar	2	0	0,032	0,026
10	biaya	2	0	0,032	0,026
11	studi	2	0	0,032	0,026
12	mahasiswa	3	0	0,043	0,026
13	mampu	1	0	0,022	0,026
14	orang	1	0	0,022	0,026
15	tua	1	0	0,022	0,026
16	dukung	1	0	0,022	0,026
17	aspek	2	0	0,032	0,026
18	adil	2	0	0,032	0,026
19	ringan	1	0	0,022	0,026
20	beban	1	0	0,022	0,026
21	ekonomi	1	0	0,022	0,026
22	prestasi	1	0	0,022	0,026
23	keluarga	1	0	0,022	0,026
24	universitas	4	0	0,054	0,026
25	brawijaya	1	0	0,022	0,026
26	wajib	1	0	0,022	0,026
27	pajadjaran	2	0	0,032	0,026
28	sebatas	1	0	0,022	0,026
29	imbau	1	0	0,022	0,026
30	rector	1	0	0,022	0,026
31	ganjar	2	0	0,032	0,026
32	kurnia	1	0	0,022	0,026
33	bandung	2	0	0,032	0,026
34	jawa	1	0	0,022	0,026
35	kemarin	1	0	0,022	0,026
36	beda	1	0	0,022	0,026
37	efektif	1	0	0,022	0,026
38	mudah	1	0	0,022	0,026

Dokumen 2 (yang sudah di preprocessing) :

(Bisnis dan Keuangan)

Judul : Inflasi Januari Capai 0,89 Persen

jakarta kompas com pusat statistik mencatat inflasi januari angka melebihi inflasi januari kepala bps rusman heriawan menilai angka inflasi januari menuding sebab harga pangan inflasi relatif masuk bulan krusial produksi panganbps mencatat januari kota inflasi kota deflasi inflasi tertinggi padang terendah manokwari

Kategori = Bisnis dan Keuangan

Daftar kata dari dokumen 2 yang diperoleh tertera pada Tabel 4 berikut :

Tabel 4. Daftar Kata Contoh Dokumen Pembelajaran 2

No.	Kata	Frekuensi	No.	Kata	Frekuensi
1	jakarta	1	16	menuding	1
2	kompas	1	17	sebab	1
3	com	1	18	harga	1
4	pusat	1	19	pangan	2
5	statistik	1	20	relative	1
6	mencatat	2	21	masuk	1
7	inflasi	6	22	bulan	1
8	januari	4	23	krusial	1
9	angka	2	24	produksi	1
10	melebihi	1	25	kota	1
11	kepala	1	26	deflasi	1
12	bps	2	27	tertinggi	1
13	rusman	1	28	padang	1
14	heriawan	1	29	terendah	1
15	menilai	1	30	manokwari	1
		Jumlah			42

Perhitungan $P(V_j)$ (1)

Diketahui : Jumlah dokumen edukasi = 1
 Jumlah dokumen bisnis keuangan = 1
 Jumlah dokumen olahraga = 1
 $P(\text{edukasi}) = 1/1 = 1$
 $P(\text{bisnis dan keuangan}) = 1/1 = 1$
 $P(\text{olahraga}) = 1/1 = 1$

Perhitungan $P(W_k | V_j)$ (2)

Diketahui : Frekuensi edukasi = 55
 Frekuensi bisnis dan keuangan = 42
 Frekuensi olahraga = 59
 Jumlah kata = 66

Pada kategori edukasi :

$P(\text{kompas} | \text{edukasi}) = (1+1) / (55+66) = 0.017$
 $P(\text{perguruan} | \text{edukasi}) = (2+1) / (55+66) = 0.025$
 $P(\text{ptn} | \text{edukasi}) = (1+1) / (55+66) = 0.017$
 $P(\text{jakarta} | \text{edukasi}) = (0+1) / (55+66) = 0.008$
 Hasil perhitungan lengkap dapat dilihat dalam Tabel 4

Pada kategori bisnis & keuangan :

$P(\text{kompas} | \text{bisnis keuangan}) = (1+1) / (42+66) = 0.019$
 $P(\text{perguruan} | \text{bisnis keuangan}) = (0+1) / (42+66) = 0.009$
 $P(\text{ptn} | \text{bisnis keuangan}) = (0+1) / (42+66) = 0.009$
 $P(\text{jakarta} | \text{bisnis keuangan}) = (1+1) / (42+66) = 0.019$
 Hasil perhitungan lengkap dapat dilihat dalam Tabel 4

Pengetahuan kata pada dokumen 2 tertera pada Tabel 5 berikut :

Tabel 5. Daftar Pengetahuan Kata Contoh Dokumen Pembelajaran 2

No	Kata	Frekuensi Edukasi	Frekuensi Bisnis & Keuangan	Probabilitas Edukasi	Probabilitas Bisnis & Keuangan
1	kompas	1	1	0,017	0,019
2	com	1	1	0,017	0,019
3	perguruan	2	0	0,025	0,009

<https://doi.org/10.31849/digitalzone.v14i1.13596>

Digital Zone is licensed under a Creative Commons Attribution International (CC BY-SA 4.0)

No	Kata	Frekuensi Edukasi	Frekuensi Bisnis & Keuangan	Probabilita s Edukasi	Probabilitas Bisnis & Keuangan
4	negeri	1	0	0,017	0,009
5	ptn	1	0	0,017	0,009
6	indonesia	2	0	0,025	0,009
7	ideal	2	0	0,025	0,009
8	tetap	2	0	0,025	0,009
9	standar	2	0	0,025	0,009
10	biaya	2	0	0,025	0,009
11	studi	2	0	0,025	0,009
12	mahasiswa	3	0	0,033	0,009
13	mampu	1	0	0,017	0,009
14	orang	1	0	0,017	0,009
15	tua	1	0	0,017	0,009
16	dukung	1	0	0,017	0,009
17	aspek	2	0	0,025	0,009
18	adil	2	0	0,025	0,009
19	ringan	1	0	0,017	0,009
20	beban	1	0	0,017	0,009
21	ekonomi	1	0	0,017	0,009
22	prestasi	1	0	0,017	0,009
23	keluarga	1	0	0,017	0,009
24	universitas	4	0	0,041	0,009
25	brawijaya	1	0	0,017	0,009
26	wajib	1	0	0,017	0,009
27	padjadjaran	2	0	0,025	0,009
28	sebatas	1	0	0,017	0,009
29	imbau	1	0	0,017	0,009
30	rector	1	0	0,017	0,009
31	ganjar	2	0	0,025	0,009
32	kurnia	1	0	0,017	0,009
33	bandung	2	0	0,025	0,009
34	jawa	1	0	0,017	0,009
35	kemarin	1	0	0,017	0,009
36	beda	1	0	0,017	0,009
37	efektif	1	0	0,017	0,009
38	mudah	1	0	0,017	0,009
39	jakarta	0	1	0,008	0,019
40	pusat	0	1	0,008	0,019
41	statistik	0	1	0,008	0,019
42	mencatat	0	2	0,008	0,028
43	inflasi	0	6	0,008	0,065
44	januari	0	4	0,008	0,046
45	angka	0	2	0,008	0,028
46	melebihi	0	1	0,008	0,019
47	kepala	0	1	0,008	0,019
48	bps	0	2	0,008	0,028
49	rusman	0	1	0,008	0,019
50	heriawan	0	1	0,008	0,019
51	menilai	0	1	0,008	0,019

No	Kata	Frekuensi Edukasi	Frekuensi Bisnis & Keuangan	Probabilitas Edukasi	Probabilitas Bisnis & Keuangan
52	menuding	0	1	0,008	0,019
53	sebab	0	1	0,008	0,019
54	harga	0	1	0,008	0,019
55	pangan	0	2	0,008	0,028
56	relative	0	1	0,008	0,019
57	masuk	0	1	0,008	0,019
58	bulan	0	1	0,008	0,019
59	krusial	0	1	0,008	0,019
60	produksi	0	1	0,008	0,019
61	kota	0	1	0,008	0,019
62	deflasi	0	1	0,008	0,019
63	tertinggi	0	1	0,008	0,019
64	padang	0	1	0,008	0,019
65	terendah	0	1	0,008	0,019
66	manokwari	0	1	0,008	0,019

Dokumen 3 (untuk testing dan sudah di preprocessing) :

(Olahraga)

Judul : Okafor Absen Tiga Bulan

new orleans kompas com center new orleans hornets emeka okafor absen cedera otot perut akhir rekor main absen tanding hornets selasa okafor rata rata poin pimpin rebound main tanding turut turut tanding runtun panjang empat nba tim tersebutn main usia tujuh nba sakit otot perut kuarter tanding lawan phoenix suns tim kalah new orleans washington selasa

Kategori = Olahraga

Daftar kata dari dokumen 3 yang diperoleh tertera pada Tabel 6 berikut :

Tabel 6. Daftar Kata Contoh Dokumen Klasifikasi

No.	Kata	Frekuensi	No.	Kata	Frekuensi
1	new	3	20	pimpin	1
2	orleans	3	21	rebound	1
3	kompas	1	22	turut	2
4	com	1	23	runtun	1
5	center	1	24	panjang	1
6	hornets	2	25	empat	1
7	emeka	1	26	nba	2
8	okafor	2	27	tim	2
9	absen	2	28	sebut	1
10	cedera	1	29	usia	1
11	otot	2	30	tujuh	1
12	perut	2	31	sakit	1
13	akhir	1	32	kuarter	1
14	rekor	1	33	lawan	1
15	main	3	34	phoenix	1
16	tanding	4	35	suns	1
17	selasa	2	36	kalah	1
18	rata	2	37	washington	1
19	poin	1			
		Jumlah	59		

Perhitungan $P(V_j)$

(1)

Diketahui : Jumlah dokumen edukasi = 1

Jumlah dokumen bisnis keuangan = 1

<https://doi.org/10.31849/digitalzone.v14i1.13596>

Digital Zone is licensed under a Creative Commons Attribution International (CC BY-SA 4.0)

$$P(\text{edukasi}) = 1/1 = 1$$

$$P(\text{bisnis dan keuangan}) = 1/1 = 1$$

Perhitungan $P(W_k | V_j)$ (2)

Diketahui : Frekuensi edukasi = 55

Frekuensi bisnis dan keuangan = 42

Frekuensi olahraga = 59

Jumlah kata = 101

Pada kategori edukasi :

$$P(\text{kompas} | \text{edukasi}) = (1+1) / (55+101) = 0.013$$

$$P(\text{perguruan} | \text{edukasi}) = (2+1) / (55+101) = 0.019$$

$$P(\text{ptn} | \text{edukasi}) = (1+1) / (55+101) = 0.013$$

$$P(\text{jakarta} | \text{edukasi}) = (0+1) / (55+101) = 0.006$$

$$P(\text{absen} | \text{edukasi}) = (0+1) / (55+101) = 0.006$$

Hasil perhitungan lengkap dapat dilihat dalam Tabel 6

Pada kategori bisnis & keuangan :

$$P(\text{kompas} | \text{bisnis keuangan}) = (1+1) / (42+101) = 0.014$$

$$P(\text{perguruan} | \text{bisnis keuangan}) = (0+1) / (42+101) = 0.007$$

$$P(\text{ptn} | \text{bisnis keuangan}) = (0+1) / (42+101) = 0.007$$

$$P(\text{jakarta} | \text{bisnis keuangan}) = (1+1) / (42+101) = 0.014$$

$$P(\text{absen} | \text{bisnis keuangan}) = (0+1) / (42+101) = 0.007$$

Hasil perhitungan lengkap dapat dilihat dalam Tabel 6

Pada kategori olahraga :

$$P(\text{kompas} | \text{olahraga}) = (1+1) / (59+101) = 0.013$$

$$P(\text{perguruan} | \text{olahraga}) = (0+1) / (59+101) = 0.006$$

$$P(\text{ptn} | \text{olahraga}) = (0+1) / (59+101) = 0.006$$

$$P(\text{jakarta} | \text{olahraga}) = (1+1) / (59+101) = 0.006$$

$$P(\text{absen} | \text{olahraga}) = (2+1) / (59+101) = 0.019$$

Hasil perhitungan lengkap dapat dilihat dalam Tabel 6

Maka pengetahuan kata yang pada dokumen 3 tertera pada Tabel 7 berikut :

Tabel 7. Daftar Pengetahuan Kata Contoh Dokumen Pembelajaran 3

No	Kata	Frek Edukasi	Frek Bisnis Keuangan	Frek Olahraga	Prob Edukasi	Prob Bisnis Keuangan	Prob Olahraga
1	kompas	1	1	1	0,013	0,014	0,013
2	com	1	1	1	0,013	0,014	0,013
3	perguruan	2	0	0	0,019	0,007	0,006
4	negeri	1	0	0	0,013	0,007	0,006
5	ptn	1	0	0	0,013	0,007	0,006
6	indonesia	2	0	0	0,019	0,007	0,006
7	ideal	2	0	0	0,019	0,007	0,006
8	tetap	2	0	0	0,019	0,007	0,006
9	standar	2	0	0	0,019	0,007	0,006
10	biaya	2	0	0	0,019	0,007	0,006
11	studi	2	0	0	0,019	0,007	0,006
12	mahasiswa	3	0	0	0,026	0,007	0,006
13	mampu	1	0	0	0,013	0,007	0,006
14	orang	1	0	0	0,013	0,007	0,006
15	tua	1	0	0	0,013	0,007	0,006

No	Kata	Frek Edukasi	Frek Bisnis Keuangan	Frek Olahraga	Prob Edukasi	Prob Bisnis Keuangan	Prob Olahraga
16	dukung	1	0	1	0,013	0,007	0,013
17	aspek	2	0	0	0,019	0,007	0,006
18	adil	2	0	0	0,019	0,007	0,006
19	ringan	1	0	0	0,013	0,007	0,006
20	beban	1	0	0	0,013	0,007	0,006
21	ekonomi	1	0	0	0,013	0,007	0,006
22	prestasi	1	0	1	0,013	0,007	0,013
23	keluarga	1	0	0	0,013	0,007	0,006
24	universitas	4	0	0	0,032	0,007	0,006
25	brawijaya	1	0	0	0,013	0,007	0,006
26	wajib	1	0	0	0,013	0,007	0,006
27	padjadjaran	2	0	0	0,019	0,007	0,006
28	sebatas	1	0	0	0,013	0,007	0,006
29	imbau	1	0	0	0,013	0,007	0,006
30	rector	1	0	0	0,013	0,007	0,006
31	ganjar	2	0	0	0,019	0,007	0,006
32	kurnia	1	0	0	0,013	0,007	0,006
33	bandung	2	0	0	0,019	0,007	0,006
34	jawa	1	0	0	0,013	0,007	0,006
35	kemarin	1	0	0	0,013	0,007	0,006
36	beda	1	0	0	0,013	0,007	0,006
37	efektif	1	0	0	0,013	0,007	0,006
38	mudah	1	0	0	0,013	0,007	0,006
39	jakarta	0	1	0	0,006	0,014	0,006
40	pusat	0	1	0	0,006	0,014	0,006
41	statistik	0	1	0	0,006	0,014	0,006
42	mencatat	0	2	0	0,006	0,021	0,006
43	inflasi	0	6	0	0,006	0,049	0,006
44	januari	0	4	0	0,006	0,035	0,006
45	angka	0	2	0	0,006	0,021	0,006
46	melebihi	0	1	0	0,006	0,014	0,006
47	kepala	0	1	0	0,006	0,014	0,006
48	bps	0	2	0	0,006	0,021	0,006
49	rusman	0	1	0	0,006	0,014	0,006
50	heriawan	0	1	0	0,006	0,014	0,006
51	menilai	0	1	0	0,006	0,014	0,006
52	menuding	0	1	0	0,006	0,014	0,006
53	sebab	0	1	0	0,006	0,014	0,006
54	harga	0	1	0	0,006	0,014	0,006
55	pangan	0	2	0	0,006	0,021	0,006
56	relative	0	1	0	0,006	0,014	0,006
57	masuk	0	1	0	0,006	0,014	0,006
58	bulan	0	1	1	0,006	0,014	0,013
59	krusial	0	1	0	0,006	0,014	0,006
60	produksi	0	1	0	0,006	0,014	0,006
61	kota	0	1	0	0,006	0,014	0,006
62	deflasi	0	1	0	0,006	0,014	0,006
63	tertinggi	0	1	0	0,006	0,014	0,006

<https://doi.org/10.31849/digitalzone.v14i1.13596>

Digital Zone is licensed under a Creative Commons Attribution International (CC BY-SA 4.0)

No	Kata	Frek Edukasi	Frek Bisnis Keuangan	Frek Olahraga	Prob Edukasi	Prob Bisnis Keuangan	Prob Olahraga
64	padang	0	1	0	0,006	0,014	0,006
65	terendah	0	1	0	0,006	0,014	0,006
66	manokwari	0	1	0	0,006	0,014	0,006
67	new	0	0	3	0,006	0,007	0,025
68	orleans	0	0	3	0,006	0,007	0,025
69	center	0	0	1	0,006	0,007	0,013
70	hornets	0	0	2	0,006	0,007	0,019
71	emeka	0	0	1	0,006	0,007	0,013
72	okafor	0	0	2	0,006	0,007	0,019
73	absen	0	0	2	0,006	0,007	0,019
74	cedera	0	0	1	0,006	0,007	0,013
75	otot	0	0	2	0,006	0,007	0,019
76	perut	0	0	2	0,006	0,007	0,019
77	akhir	0	0	1	0,006	0,007	0,013
78	rekor	0	0	1	0,006	0,007	0,013
79	main	0	0	3	0,006	0,007	0,025
80	tanding	0	0	4	0,006	0,007	0,031
81	selasa	0	0	2	0,006	0,007	0,019
82	rata	0	0	2	0,006	0,007	0,019
83	poin	0	0	1	0,006	0,007	0,013
84	pimpin	0	0	1	0,006	0,007	0,013
85	rebound	0	0	1	0,006	0,007	0,013
86	turut	0	0	2	0,006	0,007	0,019
87	runtun	0	0	1	0,006	0,007	0,013
88	panjang	0	0	1	0,006	0,007	0,013
89	empat	0	0	1	0,006	0,007	0,013
90	nba	0	0	2	0,006	0,007	0,019
91	tim	0	0	2	0,006	0,007	0,019
92	sebut	0	0	1	0,006	0,007	0,013
93	usia	0	0	1	0,006	0,007	0,013
94	tujuh	0	0	1	0,006	0,007	0,013
95	sakit	0	0	1	0,006	0,007	0,013
96	kuarter	0	0	1	0,006	0,007	0,013
97	lawan	0	0	1	0,006	0,007	0,013
98	phoenix	0	0	1	0,006	0,007	0,013
99	suns	0	0	1	0,006	0,007	0,013
100	kalah	0	0	2	0,006	0,007	0,019
101	washington	0	0	1	0,006	0,007	0,013
Jumlah					1,000	1,000	1,019

Dokumen 4 (untuk testing dan sudah di preprocessing) :**Judul : Diultimatum Durant, Pemilik Nets Tegaskan Dukungan untuk Nash**

brooklyn nets joe tsai dukung pelatih steve nash general manager sean marks lapor ultimatum kevin durant pilih mereka laporan deskripsi temu london tolak tahan tegas posisi lewat unggah twitter putus duga mantap niat hengkang minta tukar juni kata diri tidak percaya arah club sapu boston celtics putaran pertama playoff musim juara nba tengah masuk kontrak tahun juta buku rekor menang kalah james harden philadelphia kyrie irving jarang main vaksin covid

Tabel 8. Daftar Kata Contoh Dokumen Klasifikasi

No.	Kata	Frekuensi	No.	Kata	Frekuensi
1	brooklyn	1	27	Lewat	1
2	nets	3	28	Unggah	1
3	joe	1	29	Twitter	1
4	tsai	5	30	putus	1
5	dukung	1	31	duga	1
6	pelatih	1	32	mantap	1
7	steve	1	33	niat	1
8	nash	2	34	hengkang	1
9	general	1	35	minta	1
10	manager	1	36	tukar	1
11	sean	1	37	juni	1
12	marks	1	37	kata	1
13	lapor	2	38	diri	1
14	ultimatum	2	39	tidak	1
15	kevin	1	40	percaya	1
16	durant	6	41	arah	1
17	pilih	1	42	club	1
18	mereka	1	43	sapu	1
19	laporan	1	44	boston	1
20	deskripsi	1	45	celtics	1
21	temu	1	46	putaran	1
22	london	1	47	pertama	1
23	tolak	1	48	playoff	1
24	Tahan	1	49	musim	1
25	Tegas	1	50	juara	1
26	Posisi	1	51	nba	1
		Jumlah			66

Proses pengklasifikasian dengan *algoritma naïve bayes* melalui tahapan *text mining* dan *pattern discovery* yaitu berdasarkan pada probabilitas tertinggi data latih pada kategori yang sesuai. Berdasarkan hasil perhitungan probabilitas masing-masing dokumen sesuai kategori yang terdapat pada daftar pengetahuan kata contoh dokumen pembelajaran, probabilitas kategori edukasi yaitu 1,000, probabilitas kategori bisnis keuangan yaitu 1,000 dan probabilitas kategori olahraga yaitu 1,019. Berikut tabel perbandingan nilai probabilitas dari dokumen pembelajaran masing-masing kategori:

Tabel 9. Perbandingan Nilai Probabilitas Dokumen Pembelajaran Berdasarkan Kategori

Kategori	Nilai Probabilitas
Edukasi	1,000
Bisnis dan Keuangan	1,000
Olahraga	1,019

Dari tabel di atas, maka disimpulkan bahwa **dokumen 4** yang menjadi data testing termasuk ke dalam **kategori olahraga** dengan probabilitas tertinggi (1,019) dibandingkan dengan nilai probabilitas kategori edukasi dan kategori bisnis keuangan.

Pada penelitian ini, digunakan 300 dokumen berita online yang dibagi untuk data training dan data testing. Pengujian terhadap hasil dari metode klasifikasi menggunakan Skema *K-fold Cross Validation*, dimana dataset yang digunakan untuk tahapan pengujian dibagi secara acak menjadi beberapa bagian. *Fold K-1*, bagian ke 1 menjadi data uji dan sisanya menjadi data

<https://doi.org/10.31849/digitalzone.v14i1.13596>

Digital Zone is licensed under a Creative Commons Attribution International (CC BY-SA 4.0)

latih, demikian seterusnya hingga sampai *fold* K-4. Evaluasi menggunakan *confusion matrix* melibatkan perhitungan *accuracy*, *precision*, *recall* dan *error rate*, dimana dataset dibagi menjadi 4-*fold* dan masing-masing *fold* memiliki 3 data dengan persentase sebesar 75% untuk data training dan 25% untuk data uji. Nilai keseluruhan dari hasil pengujian pada semua *fold* ditunjukkan pada Tabel 10 berikut :

Tabel 10. Nilai Hasil Pengujian

Dataset	Accuracy	Precision	Recall	Error Rate
K-1	67,1 %	62%	58,9 %	43,1%
K-2	69,4 %	75%	80%	42,8%
K-3	79,6 %	75%	78%	42,8%
K-4	82,7%	76,9%	87,9%	35,4%
Rata-rata	74,7 %	72,2 %	76,2 %	41,02 %

Berdasarkan Tabel 10 di atas, dapat dilihat hasil pengujian *accuracy*, *precision* dan *recall* yang telah dilakukan menggunakan 4 data set menghasilkan rata-rata nilai *accuracy* sebesar 74,7%, *precision* 72,2%, *recall* sebesar 76,2% dan *error rate* sebesar 41,02%. Dari hasil pengujian tersebut, terlihat persentase tingkat *accuracy* yang cenderung meningkat dan persentase tingkat *error* yang kecil yaitu dibawah 50% dibandingkan dengan penelitian sebelumnya [15] yang juga menerapkan algoritma Naïve Bayes Classifier untuk klasifikasi berita dengan seleksi fitur dan boosting, didapat nilai akurasi sebesar 72%, presisi 72,3 %, dan recall 72.3 %. Hasil penelitian ini juga terjadi peningkatan nilai akurasi jika dibandingkan dengan penelitian lain yang menggunakan algoritma naïve bayes untuk melihat analisis sentimen dari review pengunjung wisata TMII berbasis web dengan mengklasifikasikan kalimat-kalimat komentar dari twiter dengan hasil uji akurasi yaitu 70%, precision 58%, dan recall 1% [16].

4. Kesimpulan

Berdasarkan penelitian yang dilakukan, Algoritma Naive Bayes telah terbukti efektif untuk klasifikasi otomatis sebuah dokumen teks dengan tahapan *text minig* (*preprocessing*, *stemming*, dan *filtering*) yang dilanjutkan dengan tahapan *pattern discovery* (*learn naïve bayes* dan *classify naïve bayes*) dengan rata-rata nilai *accuracy* sebesar 74,7%, *precision* 72,2%, *recall* sebesar 76,2% dan *error rate* sebesar 41,02% . Proses yang sederhana serta kecepatannya yang tinggi dalam tahapan pelatihan membuat Algoritma *naïve bayes* menarik untuk digunakan dalam proses klasifikasi.

Selain itu, untuk menentukan tingkat akurasi dari hasil klasifikasi, jumlah data *training* dan data *testing* yang digunakan mempunyai pengaruh terhadap tingkat akurasi hasil. Seperti halnya algoritma lain, hasil klasifikasi dengan algoritma Naïve Bayes ini belumlah sempurna, sehingga kelebihan dan kekurangan dari algoritma ini dapat dijadikan dasar penelitian selanjutnya untuk pengembangan metode klasifikasi lain, sehingga diperoleh metode yang dapat bekerja lebih efektif dan efisien, serta mengurangi kelemahan dalam proses klasifikasi yang ada saat ini.

Daftar Pustaka

- [1] Puspitasari, N., Pratama, F. et. all. 'Quality Classification of Palm Oil Varieties Using Naive Bayes Classifier', *Digital Zone : Jurnal Teknologi Informasi & Komunikasi*. 2022
- [2] Chen, P. and Yu, L. 'Use of Data Mining Technologies in an English Online Test Results Management System', *International Journal of Emerging Technologies in Learning*, 16(9), pp. 166–181. 2021 Available at: <https://doi.org/10.3991/ijet.v16i09.22743>.

- [3] Hasan, N.F. *et al.* ‘Sentiment Analysis of Public Opinion Regarding Papuan Local Languages Condition Using Data Science Approach’, *Digital Zone : Jurnal Teknologi Informasi & Komunikasi*, 13(02), pp. 125–139. 2022
- [4] Irmanita, R., Sri Suryani Prasetyowati and Yuliant Sibaroni. ‘Classification of Malaria Complication Using CART (Classification and Regression Tree) and Naïve Bayes’, *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 5(1), pp. 10–16. 2021 Available at: <https://doi.org/10.29207/resti.v5i1.2770>.
- [5] Safitri, S.N., Haryono Setiadi and Suryani, E. (2022) ‘Educational Data Mining Using Cluster Analysis Methods and Decision Trees based on Log Mining’, *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 6(3), pp. 448–456. Available at: <https://doi.org/10.29207/resti.v6i3.3935>.
- [6] Sari, R. and Hayuningtyas, R.Y. (2019) ‘Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Pada Wisata TMII Berbasis Website’, *Indonesian Journal on Software Engineering (IJSE)*, 5(2), pp. 51–60. Available at: <https://doi.org/10.31294/ijse.v5i2.6957>.
- [7] Seemi, F. *et al.* ‘Browsing behaviour analysis using data mining’, *International Journal of Advanced Computer Science and Applications*, 10(2), pp. 490–498. 2019 Available at: <https://doi.org/10.14569/ijacsa.2019.0100263>.
- [8] Sravani, T., Madala, S.R. and HeenaKausar, S. ‘College students’ Network behavior Using data mining and feature analysis’, *Journal of Physics: Conference Series*, 2089(1). 2021 Available at: <https://doi.org/10.1088/1742-6596/2089/1/012075>.
- [9] Suharjo, B. ‘Application of K-Means Cluster and Spatial Statistics using Python to Analyze the Indicators of Indonesia Information Technology’, *Digital Zone: Jurnal Teknologi Informasi dan Komunikasi*, 12(1), pp. 11–18. 2021 Available at: <https://doi.org/10.31849/digitalzone.v12i1.4310>.
- [10] Sunday, K. *et al.* ‘Analyzing student performance in programming education using classification techniques’, *International Journal of Emerging Technologies in Learning*, 15(2), pp. 127–144. 2020 Available at: <https://doi.org/10.3991/ijet.v15i02.11527>.
- [11] Sutoyo, E. and Almaarif, A. ‘Educational Data Mining for Predicting Student Graduation Using the Naïve Bayes Classifier Algorithm’, *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 4(1), pp. 95–101. 2020 Available at: <https://doi.org/10.29207/resti.v4i1.1502>.
- [12] Tope-Oke, A., Afolalu, C.A. and Omofade, O. ‘A Data Mining Based Approach to Customer Behaviour in an Electronic Settings’, *Journal of Computer and Communications*, 07(05), pp. 42–53. 2019 Available at: <https://doi.org/10.4236/jcc.2019.75004>.
- [13] Wang, C. ‘Analysis of Students’ Behavior in English Online Education Based on Data Mining’, *Mobile Information Systems*, 2021. Available at: <https://doi.org/10.1155/2021/1856690>.
- [14] Wiza, F. ‘Klasterisasi karakteristik kekerasan seksual terhadap anak dengan metode k-means cluster analysis’, *Digital Zone: Jurnal Teknologi Informasi dan Komunikasi*, 10(1), pp. 44–53. 2019 Available at: <https://doi.org/10.31849/digitalzone.v10i1.2423>.
- [15] A. Ridwan, “Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus,” *J. SISKOM-KB*, vol. 4, no. 1, pp. 15–21, 2020.
- [16] Prakoso, B.S. *et al.* ‘Klasifikasi Berita Menggunakan Algoritma Naive Bayes Classifier Dengan Seleksi Fitur Dan Boosting’, *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 3(2), pp. 227–232. 2022 Available at: <https://doi.org/10.29207/resti.v3i2.1042>.