Jurnal Ilmiah Teknologi dan Informasia ASIA (JITIKA)
Vol.10, No.1, Februari 2016
ISSN: 0852-730X

Klasifikasi Berita Lokal Radar Malang Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram

Denny Nathaniel Chandra[1], Gede Indrawan[2,] I Nyoman Sukajaya[3]
[1,2,3]Program Studi Ilmu Komputer, Program Pascasarjana
Universitas Pendidikan Ganesha
Singaraja, Indonesia
e-mail: [1]dennyn_c@yahoo.com, [2]gede.indrawan@gmail.com, [3]nyoman.sukajaya@undiksha.ac.id

ABSTRAK. Perkembangan berita digital telah tumbuh sangat cepat. Saat ini diperkirakan 80% berita digital dalam bentuk tidak terstruktur. Tingginya volume dokumen teks ini dipicu oleh aktivitas dari berbagai sumber berita. Kebutuhan analisis text mining sangat diperlukan dalam menangani teks yang tidak terstruktur tersebut. Untuk mengklasifikasikan berita, banyak peneliti yang berusaha untuk melakukan klasifikasi terhadap berita ini secara otomatis, salah satunya adalah dengan menggunakan klasifikasi naïve bayes. Pada penelitian ini selain menggunakan naïve bayes, peneliti juga akan menggunakan fitur N-Gram. Diharapkan dengan penambahan metode ini, dapat meningkatkan tingkat akurasi dari klasifikasi naïve bayes.

***Kata kunci**: klasifikasi, naive bayes classifier, akurat*

1. PENDAHULUAN

Informasi menjadi kebutuhan pokok bagi setiap orang, namun tidak semua informasi yang ada dapat menjadi kebutuhan. Dipengaruhi oleh kemajuan teknologi internet sehingga informasi mengalami pelonjakan yang besar, sementara volume berita elektronik berbahasa Indonesia yang semakin besar adalah sumber informasi yang berharga, dan memungkinkan banyak pengguna informasi untuk merubah, memperbanyak, dan menghasilkan informasi baru. Sehingga dewasa ini perlu pencermatan lebih agar mendapatkan informasi yang relevan dan sesuai dengan apa yang diinginkan oleh pengguna informasi, pengelompokan berita dibutuhkan untuk mempermudah pencarian informasi mengenai suatu *event* tertentu. Berbagai penelitian yang dilakukan oleh peneliti terdahulu mengenai text mining merupakan bukti banyaknya informasi media elektronik yang mengharuskan adanya pengembangan tentang proses penyaringan informasi secara berkala untuk menghasilkan informasi yang baik, serta dipengaruhi oleh permasalahan klasifikasi dokumen yang mendasar dan sangat penting. Dalam dokumen teks, tulisan yang terkandung adalah bahasa alami manusia, yang merupakan bahasa dengan struktur kompleks dan jumlah kata yang sangan banyak.

Salah satunya penelitian terhadap situs berita radar malang yang penulis lakukan yang bertujuan untuk mengklasifikasikan jenis berita yang sesuai dengan konten berita pada situs tersebut dan diharapkan mampu membantu menjadikan acuan bagi developer portal berita online agar dapat memenajemen konten-konten yang terdapat di dalamnya dengan baik Pada penelitian ini penulis menggunakan metode *Naïve Bayes Classification*, penelitian ini berusaha untuk mengklasifikasikan dokumen dengan metode tersebut. Klasifikasi ini ditekankan untuk dokumen berbahasa Indonesia, sementara keterkaitan antar dokumen diukur berdasarkan probabilitas.

Erfian Junianto dengan judul "Penerapan Particle Swarm Optimization Untuk Seleksi Fitur Pada Klasifikasi Dokumen Berita Menggunakan Naive Bayes Classifier". Penelitian terkait selanjutnya oleh Amir Hamzah , melakukan penelitian dengan judul "Klasifikasi Teks Dengan Naive Bayes Classifier (NBC) Untuk Pengelompokan Teks Berita dan *Abstract* Akademis*"*dan Akhmad Pandhu Wijawa , melakukan penelitian dengan judul "Klasifikasi Dokumen dengan Naïve Bayes Classification(NBC) Untuk Mengetrahui Konten E-Goverment". Melihat hasil dari penelitian tersebut menjadikan Naïve Bayes Classifier sebagai metode yang dipilih pada penelitian ini.

Penggunaan Naïve Bayes Classifier pada penelitian ini diharapkan mampu menghasilkan klasifikasi yang akurat agar dapat dijadikan bahan penelitian lebih lanjut, kelebihan Naïve Bayes Classifier dibandingkan algoritma lain adalah pada kemampuannya mengklasifikasi dokumen dengan kesederhanaan dan kecepatan komputasinya namun memiliki komputasi tinggi, metode Naïve Bayes Classifier juga memiliki kinerja yang baik terhadap pengklasifikasian data dokumen yang mengandung angka maupun teks.

2. METODE YANG DIUSULKAN

Text mining adalah cara yang digunakan untuk ekstraksi informasi yang lebih berkualitas dari dataset yang tersedia. Penelitian ini mengusulkan metode klasifikasi dengan algoritma *Naïve Bayes Classification* (NBC).

a. Text mining

Penambangan teks (bahasa Inggris: text mining) adalah proses ekstraksi pola berupa informasi dan pengetahuan yang berguna dari sejumlah besar sumber data teks, seperti dokumen Word, PDF, kutipanteks, dll. Jenis masukan untuk penambangan teksini disebut data takter struktur dan merupakan pembeda utama dengan penambangan data yang menggunakan data terstruktur atau basis data sebagai masukan[6]. Penambangan teks dapat dianggap sebagai proses dua tahap yang diawali dengan penerapan struktur terhadap sumber data teks dan dilanjutkan dengan ekstraksi informasi dan pengetahuan yang relevandari data teks terstrukturini dengan menggunakan teknik dan alat yang sama dengan penambangan data. Proses yang umum dilakukan oleh penambangan teks di antaranya adalah perangkuman otomatis, kategorisasi dokumen, penggugusan teks, dan lain-lain.

b. Berita

Berita merupakan bentuk laporan tentang suatu kejadian yang sedang terjadi baru baru ini atau keterangan terbaru dari suatu peristiwa. Dengan kata lain berita adalah fakta menarik atau sesuatu hal yang penting yang disampaikan pada masyarakat orang banyak melalui media. Tapi tidak semua fakta bisa diangkat menjadi suatu berita oleh media. Karena setiap fakta akan dipilih mana yang pantas untuk disampaikan pada masyarakat.

c. *Preprocessing*

Pada *text preprocessing*, terdapat beberapa langkah seperti case folding, *tokenizing, filtering, dan stemming*.

a. Case Folding

Case Folding mengubah semua huruf dalam dokumen menjadi huruf kecil. Karakter selain huruf dihilangkan dan dianggap delimeter.

b. Tokenisasi

Tokenisasi adalah tugas memisahkan deretan kata didalam kalimat, paragraf atau halaman menjadi token atau potongan kata tunggal atau *termmed word.*

c. Filtering

Tahap filtering adalah tahap mengambil kata-kata penting dari hasil token. Bisa menggunakan algoritma *stoplist* (membuang kata-kata yang kurang penting) atau *wordlist* (menyimpan kata-kata penting). Contoh *stopwords* adalah "yang", "dan", "di", "dari" dan seterusnya.

d. Stemming

Stemming adalah tahap mencari *root* kata / frase dari hasil filtering. Pada tahap ini dilakukan proses pengembalian berbagai bentuk kata kedalam suatu representasi yang sama.

d. Fitur N-gram

N-gram adalah *substring* sepanjang *n* karakter dari sebuah *string* dalam definisi lain *n-gram* adalah potongan sejumlah *n* karakter dari sebuah *string*. *N-gram* merupakan sebuah metode yang diaplikasikan untuk pembangkitan kata atau karakter. Metode *n-gram* ini digunakan untuk mengambil potongan-potongan karakter huruf sejumlah *n* dari sebuah kata yang secara kontinuitas dibaca dari teks sumber hingga akhir dari dokumen.

N-gram dibedakan berdasarkan jumlah potongan karakter sebesar *n*. Untuk membantu dalam mengambil potongan-potongan kata berupa karakter huruf tersebut, maka dilakukan *padding* dengan *blank* diawal dan diakhir suatu kata. Sebagai contoh : kata "TEXT" dapat diuraikan ke dalam beberapa *n-gram* berikut ("_" merepresentasikan *blank*) :

uni-grams : T, E, X, T
bi-grams : _T, TE, EX, XT, T_
tri-grams : _TE, TEX, EXT, XT_
quad-grams : _TEX, TEXT, EXT_
quint-grams : _TEXT, TEXT_

Pembentukan Model N-gram Dokumen

Model *n-gram* dokumen dibentuk berdasarkan frekuensi *n-gram* yang muncul di dalam dokumen. Dokumen akan dibaca kata per kata, dan untuk setiap kata akan dibuat *n-gram* dari kata tersebut. Untuk setiap *n-gram* yang dibangkitkan, akan dicatat dalam sebuah *table* dengan *n-gram* sebagai kunci dan jumlah sebagai isi. Apabila *n-gram* tersebut sudah pernah muncul di dalam dokumen maka frekuensi untuk *n-gram* itu akan ditambah satu, jika belum maka *n-gram* tersebut akan ditambahkan ke dalam *table* dengan jumlah kemunculan satu.

Sebagai contoh untuk pembentukan model *n-gram* dokumen yang menggunakan *bi-gram* pada sebuah dokumen yang hanya berisi satu kalimat, "*pengenalan bahasa suku bangsa indonesia berbasis teks dengan menggunakan metode ngram*". Akan menghasilkan *bi-gram* (*2-gram*) pada Tabel 1.

Tabel 1. Contoh pembentukan N-gram

No	Kata	*Bi-gram*
1	pengenalan	_p; pe; en; ng; ge; en; na; al; la; an; n_
2	bahasa	_b; ba; ah; ha; as; sa; a_
3	suku	_s; su; uk; ku; u_
4	bangsa	_b; ba; an; ng; gs; sa; a_
5	indonesia	_i; in; nd; do; on; ne; es; si; ia; a_
6	berbasis	_b; be; er; rb; ba; as; si; is; s_
7	teks	_t; te; ek; ks; s_
8	dengan	_d; de; en; ng; ga; an; n_
9	menggunakan	_m; me; en; ng; gg; gu; un; na; ak; ka; an; n_
10	metode	_m; me; et; to; od; de; e_
11	ngram	_n; ng; gr; ra; am; m_

e. Naive Bayes Clasifier

Naive bayes classifier merupakan sebuah metode klasifikasi yang berakar pada teorema bayes. Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes. Ciri utama dari naive bayes classifier ini adalah asumsi yang sangat kuat (naif) akan independensi dari masing-masing kondisi atau kejadian. Dasar dari teorema naive bayes yang digunakan dalam pemrograman adalah rumus Bayes berikut ini:

$$P(A|B) = \big(P(A|B) * P(A)\big)/P(B)$$

Peluang kejadian A sebagai B ditentukan dari peluang B saat A, peluang A, dan peluang B. Pada pengaplikasiannya nanti rumus ini berubah menjadi :

$$P(C_i|D) = \big(P(D|C_i) * P(C_i)\big)/P(D)$$

Naive Bayes Classifier merupakan model penyederhanaan dari algoritma bayes yang cocok dalam pengklasifikasian text atau dokumen. Persamaannya adalah :

$$V_{MAP} = \arg\max P\big(v_j|a_1, a_2, \dots., a_n\big)$$

Berdasarkan persamaan ini, maka rumus bayes dapat ditulis menjadi :

$$V_{MAP} = \underset{v_j \in V}{\arg max}\ \frac{P\big(a_1, a_2 \dots a_n|v_j\big)P\big(v_j\big)}{P(a_1, a_2 \dots a_n)}$$

$P(a_1, a_2, \dots., a_n)$ merupakan bilangan konstan, sehingga dapat dihilangkan menjadi

$$V_{MAP} = \underset{v_j \in V}{\arg max}\ P\big(a_1, a_2 \dots a_n|v_j\big)P\big(v_j\big)$$

Karena $P\big(a_1, a_2, \dots., a_n|v_j\big)$ sulit untuk dihitung, maka akan diasumsikan bahwa setiap kata pada dokumen tidak mempunyai keterkaitan

$$V_{MAP} = \underset{v_j \in V}{\arg\ max}\ P\big(v_j\big)\prod\nolimits_i P\big(a_i|v_j\big)$$

Keterangan :

$$P\big(v_j\big) = \frac{|docs_j|}{|Contoh|}$$

$$P\big(w_k|v_j\big) = \frac{n_k + 1}{n + |kosakata|}$$

Dimana untuk :
$P(v_j)$ adalah probabilitas setiap dokumen terhadap sekumpulan dokumen.
$P(w_k|v_j)$ adalah probabilitas kemunculan kata w_k pada suatu dokumen dengan kategori kelas v_j
$|docs|$ adalah frekuensi dokumen pada setiap kategori.
$|Contoh|$ adalah jumlah dokumen yang ada.
n_k adalah frekuensi kata ke-k pada setiap kategori kosakata adalah jumlah kata pada dokumen test.

3. IMPLEMENTASI

Pada bab ini akan dibahas mengenai penjelasan langkah-langkah dalam persiapan dokumen, tahap ini meliputi *converting* dan *filtering* kemudian pemrosesan file seperti proses pengenalan pola klasifikasi, metode pengukuran dan hasil pengukuran, kualitas informasi pada klasifikasi dokumen menggunakan metode *Naïve Bayes Classification*.

A. Dokumen yang Digunakan

Metode naïve bayes ckassification menempuh dua tahap dalam proses klasifikasi teks, yaitu tahap pelatihan dan tahap klasifikasi. Pada tahap pelatihan dilakukan proses analisis terhadap sampel dokumen berupa pemilihan vocabulary, yaitu kata yang mungkin muncul dalam koleksi dokumen sampel yang sedapat mungkin merepresentasikan dokumen, pada tahap pelatihan terdapat dokumen *training* yang menjadi acuan untuk proses *testing*.

1. Dokumen *training*

Berfungsi untuk pembentukan kelas dan sebagai acuan bagaimana dokumen akan diklasifikasikan, dalam penelitian ini penulis menggunakan sumber data yang telah diklasifikasikan menjadi dokumen politik,ekonomi, news, edukasi, kesehatan, travel, dan olahraga pada portal www.kompas.com.

2. Dokumen *testing*

Dalam penelitian yang dilakukan, jenis dokumen yang digunakan dalam penelitian ini diambil pada portal www.radarmalang.co.id. Data yang diambil merupakan berita-berita yang terdapat pada website tersebut yang selanjutnya akan di klasifikasikan dengan acuan dokumen training.

B. *Preprocessing* Dokumen

Tahap yang dilakukan sebelum proses klasifikasi adalah *preprocessing* untuk mencari makna pada dokumen *training* maupun *testing* dan mendukung proses klasifikasi, proses ini harus dilakukan karena pada data uji dokumen berupa paragraf beserta tag-tag yang menghilangkan arti dari dokumen tersebut. Penulis mengalami kesulitan dalam memahami isi dokumen uji sebelum dilakukan proses *preprocessing. Preprocessing* juga dapat mempengaruhi identifikasi teks yang bertujuan menentukan fitur. Hal pertama dalam pemrosesan dokumen adalah memecah kumpulan karakter ke dalam kata atau token, sering disebut sebagai tokenisasi. Tokenisasi adalah hal yang kompleks untuk program komputer karena beberapa karakter dapat ditemukan sebagai *token delimiters.* Delimiter adalah karakter spasi, tab, dan baris, sedangkan karakter ( ) < > ! ? “ kadangkala dijadikan delimiter namun tergantung pada lingkungannya.(D. Xhemali,dkk, 2009)

C. Proses Identifikasi

Proses identifikasi teks sangatlah penting untuk mengenali pola teks yang akan diklasifikasikan dan mengenali jenis – jenis teks yang akan digunakan sebagai *training.* Permasalahan yang timbul saat identifikasi adalah tidak teraturnya pola teks yang didapatkan meski telah diproses menggunakan *stopwords* pada langkah

Pada proses identifikasi yang dilakukan penulis perlu membuka dokumen satu persatu untuk memahami pola yang ada pada teks tersebut, untuk pola sendiri didapatkan tidak beraturan dalam peletakan konten.

4. HASIL DAN PEMBAHASAN

Percobaan 1

Ujicoba pertama dilakukan dengan memilih secara acak data latih dan data uji dari 764 data berita, 7 kategori. Data latih yang dipilih sebanyak 61.65% dari masing-masing kategori yang ada dan sisanya 38.35% dijadikan data uji. Jumlah berita pada data latih untuk percobaan ini adalah 471 berita dan data uji sebanyak 293 berita. Komposisi pembagian data uji dan data latih untuk percobaan ini adalah pada Tabel 2.

Tabel 2. Komposisi Data Latih dan Uji Percobaan 1

Kategori	Training	Testing
ekonomi	69	36
News	119	169
Edukasi	50	7
Kesehatan	50	9
Olahraga	50	45
Entertainment	64	7
Dan lain-lain	69	12

Melalui percobaan ini didapatkan ketepatan prediksi atau akurasi sebesar 78.66% (diprediksi secara tepat sebanyak 601 dari 764 data berita uji), dan error sebesar 17.80% (diprediksi kurang tepat sebanyak 163 dari 764 data berita uji). Berikut ini pada Tabel 3 merupakan *confusion matrix* dari ujicoba yang dilakukan:

Tabel 3. Confusion Matrix Percobaan 1

Actual/Predicted	Ekonomi	News	Edukasi	Kesehatan	Olahraga	Entertainment	Dan lain-lain
Ekonomi	31	0	0	0	0	0	0
News	100	0	0	0	0	0	0
Edukasi	6	0	0	0	0	0	0
Kesehatan	6	0	0	0	0	0	0
Olahraga	31	3	0	0	0	1	0
Entertainment	7	0	0	0	0	0	0
Dan lain-lain	9	0	0	0	0	0	1

Percobaan 2

Ujicoba kedua dilakukan dengan memilih secara acak data latih dan data uji dari 610 data berita, 7 kategori. Data latih yang dipilih sebanyak 63.11% dari masing-masing kategori yang ada dan sisanya 36.89% dijadikan data uji. Jumlah berita pada data latih untuk percobaan ini adalah 385 berita dan data uji sebanyak 225 berita. Komposisi pembagian data uji dan data latih untuk percobaan ini seperti pada Tabel 4.

Tabel 4. Komposisi Data Latih dan Uji Percobaan 2

Kategori	Training	Testing
ekonomi	55	30
News	100	125
Edukasi	40	7
Kesehatan	40	9
Olahraga	40	35
Entertainment	50	7
Dan lain-lain	60	12

Melalui percobaan ini didapatkan ketepatan prediksi atau akurasi sebesar 68.20% (diprediksi secara tepat sebanyak 416 dari 610 data berita uji), dan error sebesar 31.80% (diprediksi kurang tepat sebanyak 194 dari 610 data berita uji). Tabel 5 merupakan *confusion matrix* dari ujicoba yang dilakukan.

Tabel 5. Confusion Matrix Percobaan 2

Actual/Predicted	Ekonomi	News	Edukasi	Kesehatan	Olahraga	Entertainment	Dan lain-lain
Ekonomi	30	0	0	0	0	0	0
News	125	0	0	0	0	0	0
Edukasi	7	0	0	0	0	0	0
Kesehatan	9	0	0	0	0	0	0
Olahraga	32	0	0	0	0	3	0
Entertainment	7	0	0	0	0	0	0
Dan lain-lain	11	0	0	0	0	0	1

Percobaan 3

Ujicoba ketiga dilakukan dengan memilih secara acak data latih dan data uji dari 314 data berita, 7 kategori. Data latih yang dipilih sebanyak 57.32% dari masing-masing kategori yang ada dan sisanya 42.68% dijadikan data uji. Jumlah berita pada data latih untuk percobaan ini adalah 180 berita dan data uji sebanyak 134 berita. Komposisi pembagian data uji dan data latih untuk percobaan ini adalahbisa dilihat pada Tabel 6.

Tabel 6. Komposisi Data Latih dan Uji Percobaan 3

Kategori	Training	Testing
ekonomi	40	20
News	60	75
Edukasi	25	7
Kesehatan	25	7
Olahraga	30	25
Entertainment	40	5
Dan lain-lain	43	10

Melalui percobaan ini didapatkan ketepatan prediksi atau akurasi sebesar 59.24% (diprediksi secara tepat sebanyak 186 dari 314 data berita uji), dan error sebesar 40.76% (diprediksi kurang tepat sebanyak 128 dari 314 data berita uji). C*onfusion matrix* dari ujicoba yang dilakukan ditunjukkan pada Tabel 7.

Tabel 7. Confusion Matrix Percobaan 3

Actual/Predicted	Ekonomi	News	Edukasi	Kesehatan	Olahraga	Entertainment	Dan lain-lain
Ekonomi	20	0	0	0	0	0	0
News	75	0	0	0	0	0	0
Edukasi	7	0	0	0	0	0	0
Kesehatan	7	0	0	0	0	0	0
Olahraga	24	0	0	0	0	1	0
Entertainment	5	0	0	0	0	0	0
Dan lain-lain	9	0	0	0	0	0	1

Percobaan 4

Ujicoba keempat dilakukan dengan memilih secara acak data latih dan data uji dari 361 data berita, 7 kategori. Data latih yang dipilih sebanyak 67.87% dari masing-masing kategori yang ada dan sisanya 32.13% dijadikan data uji. Jumlah berita pada data latih untuk percobaan ini adalah 245 berita dan data uji sebanyak 116 berita. Komposisi pembagian data uji dan data latih untuk percobaan ini ditunjukkan pada Tabel 8.

Tabel 8. Komposisi Data Latih dan Uji Percobaan 4

Kategori	Training	Testing
ekonomi	35	20
News	60	50
Edukasi	25	7
Kesehatan	25	7
Olahraga	30	20
Entertainment	30	5
Dan lain-lain	40	7

Melalui percobaan ini didapatkan ketepatan prediksi atau akurasi sebesar 65.93% (diprediksi secara tepat sebanyak 238 dari 361 data berita uji), dan error sebesar 34.07% (diprediksi kurang tepat sebanyak 123 dari 361 data berita uji). Berikut Tabel 9 merupakan *confusion matrix* dari ujicoba yang dilakukan:

Tabel 9. Confusion Matrix Percobaan 4

Actual/Predicted	Ekonomi	News	Edukasi	Kesehatan	Olahraga	Entertainment	Dan lain-lain
Ekonomi	30	0	0	0	0	0	0
News	60	0	0	0	0	0	0
Edukasi	7	0	0	0	0	0	0
Kesehatan	9	0	0	0	0	0	0
Olahraga	32	0	0	0	0	3	0
Entertainment	7	0	0	0	0	0	0
Dan lain-lain	11	0	0	0	0	0	1

Percobaan 5

Ujicoba kelima dilakukan dengan memilih secara acak data latih dan data uji dari 246 data berita, 7 kategori. Data latih yang dipilih sebanyak 69.10% dari masing-masing kategori yang ada dan sisanya 30.89% dijadikan data uji. Jumlah berita pada data latih untuk percobaan ini adalah 170 berita dan data uji sebanyak 76 berita. Komposisi pembagian data uji dan data latih untuk percobaan ini adalah seperti pada Tabel 10.

Tabel 10. Komposisi Data Latih dan Uji Percobaan 5

Kategori	Training	Testing
ekonomi	30	15
News	40	25
Edukasi	20	7
Kesehatan	15	7
Olahraga	20	10
Entertainment	25	5
Dan lain-lain	20	7

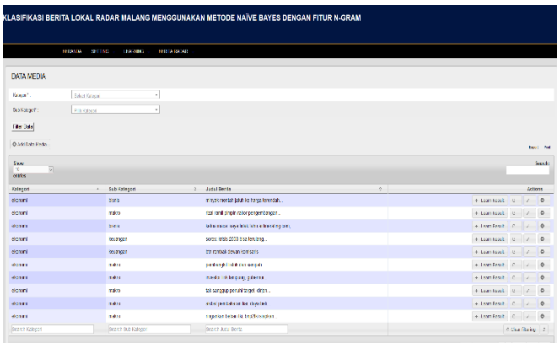
Melalui percobaan ini didapatkan ketepatan prediksi atau akurasi sebesar 74.39% (diprediksi secara tepat sebanyak 183 dari 246 data berita uji), dan error sebesar 25.61% (diprediksi kurang tepat sebanyak 63 dari 246 data berita uji). Tabel 11 berikut ini adalah *confusion matrix* dari ujicoba yang dilakukan:

Tabel 11. Confusion Matrix Percobaan 5

Actual/Predicted	Ekonomi	News	Edukasi	Kesehatan	Olahraga	Entertainment	Dan lain-lain
Ekonomi	15	0	0	0	0	0	0
News	25	0	0	0	0	0	0
Edukasi	7	0	0	0	0	0	0
Kesehatan	7	0	0	0	0	0	0
Olahraga	10	0	0	0	0	3	0
Entertainment	5	0	0	0	0	0	0
Dan lain-lain	9	0	0	0	0	0	1

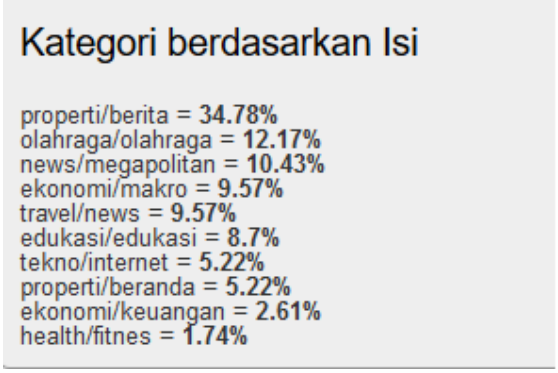
Layout Program Pengklasifikasian Berita

Gambar 1 sampai Gambar 3 Berikut ini adalah contoh tampilan proses pengklasifikasian berita online radar malang sampai dengan hasil pengklasifikasian kategori berita yang sesuai dengan ini dari salah satu berita radar malang



Gambar 1. Tampilan kumpulan berita radar malang

Gambar 2. Contoh berita



Gambar 3. hasil pengkategorian berita

Evaluasi Uji Coba

Hasil perhitungan nilai akurasi pengklasifikasian dengan Naive Bayes yang didapat dari kelima ujicoba secara berturut-turut adalah 78.66%, 68.20%, 59.24%, 65.93%, dan 74.39%. nilai akurasi terendah adalah 59.24% dan nilai akukasi tertinggi adalah 78.66%.

Melalui pemeriksaan lebih lanjut, kemungkinan akurasi dapat ditingkatkan jika data latih yang diperoleh semakin banyak, sebab klasifikasi naïve bayes merupakan metode *supervised learning* yang sangat bergantung pada data latih. Selain itu, ada beberapa kata kunci yang didapatkan dari misalnya berita news yang merupakan kata kunci juga yang dimiliki oleh berita ekonomi hal ini juga berlaku untuk kategori berita lainnya, sehingga pengklasifikasian menjadi kurang tepat dan mengindikasikan bahwa suatu berita tidak hanya mewakili satu kategori melainkan beberapa kategori.

5. KESIMPULAN DAN SARAN

Berdasarkan perancangan dan hasil analisis yang dilakukan dalam penelitian, dapat disimpulkan hal-hal berikut ini.

1. Penggunaan *N-Gram* pada penelitian ini terbukti mampu menambah jenis kata sebelum masuk ke proses stemming. Dengan banyaknya tambahan jenis kata ini sangat membantu proses klasifikasi Naïve Bayes menjadi lebih efektif dan akurat.
2. Pada penelitian klasifikasi Naïve Bayes dengan fitur *N-Gram* ini hasil akurasi maksimalnya adalah 78.66% untuk data uji berita ekonomi, news, edukasi, kesehatan, olahraga, entertainment, dan lain-lain dalam Bahasa indonesia.

Saran-saran yang muncul setelah dilakukan penelitian ini adalah sebagai berikut:

1. Penelitian ini mengklasifikasikan dokumen menggunakan *Naive Bayes Classification* dengan fitur N-Gram dalam penelitian selanjutnya dapat dikembangkan dengan metode klasifikasi lainnya seperti *Support Vector Machine, Neural Network*.
2. Memperbaiki pengolahan dan identifikasi dokumen serta mengembangkan tahap *preprocessing* dengan menyeleksi lebih banyak kata-kata yang dianggap tidak perlu ada pada dokumen untuk meningkatkan proses klasifikasi dokumen.

DAFTAR PUSTAKA

[1] A. Hamzah. 2012. *Klasifikasi Teks Dengan Naive Bayes Classifier (NBC) Untuk Pengelompokan Teks Berita dan Akademis. ISSN:1979-911I*, vol. 3.

[2] A. P. Wijaya. 2014. *Klasifikasi Dokumen dengan Naive Bayes Classification (NBC) Untuk Mengatahui Konten.* pp. 1–6.

[3] D. Xhemali, C. J. Hinde, and R. G. Stone. 2009. *Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages.* vol. 4, no. 1, pp. 16–23.

[4] E. Junianto. 2014. *Penerapan Particle Swarm Optimization Untuk Seleksi Fitur Pada Klasifikasi Dokumen Berita Menggunakan Naive Bayes Classifier*. Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri: Jakarta.