Klasifikasi Berita Berbahasa Indonesia Menggunakan *Naïve Bayes Classifier*

Qurrota 'Aini Muthmainnah, Dian Palupi Rini, Desty Rodiah
Teknik Informatika
Universitas Sriwijaya
Palembang, Indonesia
qurrotaam1605@gmail.com, dian.palupi.rini@gmail.com, destyrodiah@gmail.com

***Abstrak*— Berita pada awalnya disalurkan melalui media seperti televisi, radio dan koran, namun dengan kemajuan teknologi saat ini membuat digitalisasi informasi lebih mudah, berita berbentuk teks digital lebih cepat tersebar, aktual dan murah, sehingga dapat mengalami pelonjakan yang besar. Oleh karena itu, perlu adanya sistem yang bisa mengklasifikasikan berita secara otomatis sesuai dengan kategori-kategori berita yang ada, dengan menggunakan metode klasifikasi teks, maka kumpulan dokumen yang jumlahnya sangat besar tersebut dapat diorganisir, sehingga dapat mempermudah dan mempercepat pencarian informasi yang dibutuhkan. Dalam penelitian ini, klasifikasi teks berita menggunakan metode *Naïve Bayes Classifier* untuk mengklasifikasikan ke dalam empat kategori yaitu, bencana alam, kesehatan, olahraga dan pendidikan. Pengujian dilakukan sebanyak empat kali dengan pembagian data yang berbeda-beda, dan hasil akurasi yang didapat yaitu pengujian pertama 100%, pengujian kedua 100%, pengujian ketiga 98,33% dan pengujian keempat 96,25%. Dari hasil tersebut, dapat disimpulkan bahwa hasil klasifikasi teks berita sudah baik.**

Kata Kunci*— *Klasifikasi Teks, Naïve Bayes Classifier (NBC), Berita

I. LATAR BELAKANG

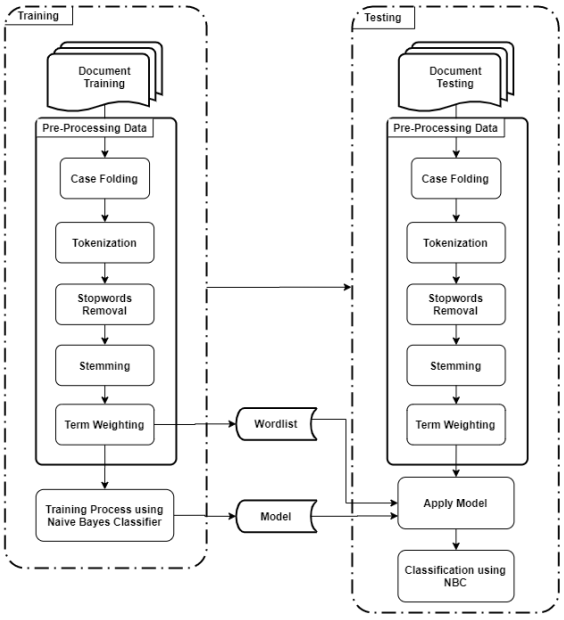
Dalam perkembangannya, pada awalnya berita disalurkan melalui media seperti televisi, radio, koran, dan sekarang dengan kehadiran personal *computer* (PC), smartphone dan perkembangan internet membuat digitalisasi informasi lebih mudah. Berita dapat tersebar lebih cepat, aktual dan murah yang berbentuk teks digital, sehingga dapat mengalami pelonjakan yang besar.

Banyaknya teks berita membuat pembaca sulit untuk menemukan jenis kategori yang diperlukan. Untuk mendapatkan jenis kategori dalam dokumen teks secara manual, pembaca harus membaca isi dokumen. Jika dokumen teks sangat panjang, maka dibutuhkan waktu yang lama bagi pembaca untuk mendapatkan jenis kategori dari sebuah berita.

Dengan menggunakan metode klasifikasi teks, maka kumpulan dokumen yang jumlahnya sangat besar tersebut dapat diorganisir, sehingga dapat mempermudah dan mempercepat pencarian jenis kategori dari sebuah berita. Pada penelitian ini, diusulkan metode *Naïve Bayes Classsifier*. NBC adalah sebuah algoritma analisa statistik, yang bekerja dengan mengolah data numerik, sering digunakan dalam penelitian tentang klasifikasi teks karena kesederhanaan dan efektivitasnya yang menggunakan ide dasar probabilitas gabungan dari kata-kata dan kategori untuk memperkirakan probabilitas kategori pada suatu dokumen [1]. Menurut [2] kelebihan dari metode NBC adalah algoritma yang sederhana dengan kompleksitas perhitungan yang rendah. Algoritma ini memiliki akurasi yang tinggi.

II. METODOLOGI PENELITIAN

Pada penelitian ini terdapat tahapan kerangka kerja yang digambarkan pada Gambar 1, yaitu pengumpulan data, *preprocessing* data, *term weighting*, dan klasifikasi dengan NBC.



Gambar 1. Kerangka Kerja Proses Perangkat Lunak

A. Data

Jenis data yang digunakan pada penelitian ini adalah data sekunder berupa teks berita berbahasa Indonesia dan berekstensi .txt. Teks berita yang digunakan ada 200 teks yang diperoleh dari enam situs media massa *online* berita nasional Indonesia. Selain data teks berita digunakan juga data berupa kamus *stopwords* bahasa Indonesia dan kamus kata dasar Bahasa Indonesia.

Data teks berita akan melalui tahap preprocessing yang terdiri dari *case folding*, *tokenization*, *stopwords removal*, dan *stemming*. Setelah itu, dilakukan *term weighting* menggunakan *term frequency*. Hasil *term weighting* akan menjadi masukan pada proses klasifikasi menggunakan NBC.

B. Naïve Bayes Classifier

Naïve Bayes Classifier (NBC) merupakan suatu metode *supervised document classification* yang menggunakan perhitungan probabilitas [3], dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Konsep dasar yang digunakan adalah Teorema *Bayes* yang dinyatakan pertama kali oleh Thomas Bayes [4], yaitu memprediksi peluang dimasa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai Teorema Bayes.

Metode ini menganggap semua atribut pada setiap kategori tidak memiliki ketergantungan satu sama lain [5]. Keuntungan penggunaan yaitu data latih untuk menentukan parameter mean dan varians dari variabel yang diperlukan untuk klasifikasi hanya sejumlah kecil [6].

NBC merupakan model penyederhanaan dari algoritma *bayes* yang cocok dalam pengklasifikasian teks atau dokumen. Persamaannya adalah :

$$V_{MAP} = \arg\max P(v_j|a_1, a_2, \dots, a_n) \quad (1)$$

Dasar dari teorema *Naive Bayes Classifier* adalah rumus *Bayes:*

$$P(A|B) = \frac{(P(A|B)xP(A))}{P(A)} \quad (2)$$

Berdasarkan persamaan (1), maka rumus *bayes* dapat ditulis menjadi :

$$V_{MAP} = \underset{v_j \in V}{\arg max} \frac{P(a_1, a_2, \dots, a_n|v_j)P(v_j)}{P(a_1,a_2,\dots,a_n)} \quad (3)$$

P(a_1,a_2,..,a_n) bilangan konstan, sehingga dapat dihilangkan menjadi :

$$V_{MAP} = \underset{v_j \in V}{\arg max} P(a_1, a_2, \dots, a_n|v_j)P(v_j) \quad (4)$$

Karena P($a1$,$a2$,….,$a_n|v_j$) sulit untuk dihitung, maka akan diasumsikan bahwa setiap kata pada dokumen tidak mempunyai keterkaitan :

$$V_{MAP} = \underset{v_j \in V}{\arg max} P(v_j) \prod_i P(a_i|v_j) \quad (5)$$

Untuk menghitung setiap kata yang terdapat pada dokumen latih dapat digunakan persamaan (6).

$$P(v_j) = \frac{(docs_j)}{(sample)} \quad (6)$$

Keterangan :

P(v_j) : probabilitas dokumen kategori v_j
($docs_j$) : jumlah seluruh dokumen pada suatu kategori
(sample) : jumlah seluruh dokumen latih

Sedangkan untuk menghitung probabilitas kategori dokumen digunakan persamaan (7).

$$P(w_k|v_j) = \frac{n_k+1}{n+|\text{vocab}|} \quad (7)$$

Keterangan :

$P(W_k|v_j)$: probabilitas kata w_k dalam kategori v_j
n_k : jumlah kemunculan kata dalam kategori
n : jumlah kata dalam kategori
|vocab| : jumlah kata dari semua kategori

III. HASIL DAN DISKUSI

Pengujian dalam penelitian ini dilakukan dengan menghitung nilai akurasi secara manual, akurasi didapatkan dengan membandingkan hasil antara jumlah klasifikasi benar oleh sistem terhadap jumlah dokumen uji. Pengujian ini dilakukan sebanyak empat kali pengujian, dengan pembagian data yang dapat dilihat pada Tabel 1.

Tabel 1 Pembagian Dara Untuk Pengujian

Pengujian	Data Latih	Data Uji	Jumlah Data (Tiap Kategori)	
			Data Latih	Data Uji
Pengujian 1	90%	10%	45	5
Pengujian 2	80%	20%	40	10
Pengujian 3	70%	30%	35	15
Pengujian 4	60%	40%	30	20

Berdasarkan hasil pengujian yang telah dilakukan maka dapat diperoleh hasil akurasi dari keempat pengujian dalam proses klasifikasi pada tiap kategori pada data uji dapat dilihat pada Tabel 2, Tabel 3, Tabel 4 dan Tabel 5.

Tabel 2. Hasil Akurasi Pengujian Pertama

Kategori	Total Teks Berita	Total Sistem	Total Benar	Akurasi
Bencana Alam	5	5	5	100%
Kesehatan	5	5	5	100%
Olahraga	5	5	5	100%
Pendidikan	5	5	5	100%
Total	**20**	**20**	**20**	**100%**

Tabel 2. Hasil Akurasi Pengujian Kedua

Kategori	Total Teks Berita	Total Sistem	Total Benar	Akurasi
Bencana Alam	10	10	10	100%
Kesehatan	10	10	10	100%
Olahraga	10	10	10	100%
Pendidikan	10	10	10	100%
Total	**40**	**40**	**40**	**100%**

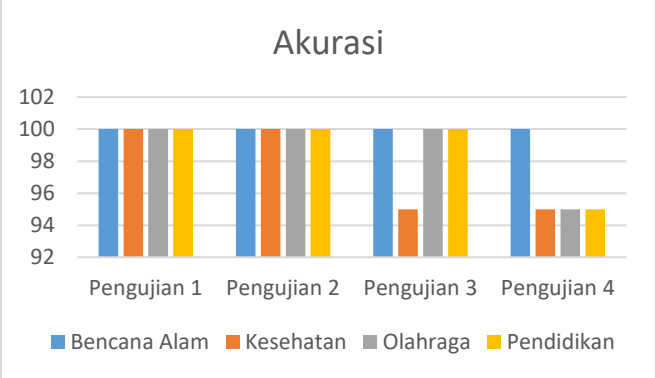
Tabel 3. Hasil Akurasi Pengujian Ketiga

Kategori	Total Teks Berita	Total Sistem	Total Benar	Akurasi
Bencana Alam	15	15	15	100%
Kesehatan	15	14	14	93,33%
Olahraga	15	15	15	100%
Pendidikan	15	15	15	100%
Total	**60**	**59**	**59**	**98,33%**

Tabel 4. Hasil Akurasi Pengujian Keempat

Kategori	Total Teks Berita	Total Sistem	Total Benar	Akurasi
Bencana Alam	20	20	20	100%
Kesehatan	20	19	19	95%
Olahraga	20	19	19	95%
Pendidikan	20	19	19	95%
Total	**80**	**77**	**77**	**96,25%**

Gambar 2 menunjukkan perbandingan nilai akurasi dari hasil proses klasifikasi menggunakan *Naïve Bayes Classifier* pada tiap kategori klasifikasi dalam bentuk grafik batang.



Gambar 2. Perbandingan Nilai Akurasi Klasifikasi Teks Berita

Berdasarkan Gambar 2, dapat dilihat nilai akurasi paling tinggi ada pada pengujian pertama dan kedua dengan rata-rata akurasi 100%. Sementara itu, pengujian ketiga klasifikasi kategori bencana alam, kesehatan, olahraga dan pendidikan menghasilkan nilai akurasi rata-rata sebesar 98,33%. Pengujian keempat klasifikasi kategori bencana alam, kesehatan, olahraga dan pendidikan menghasilkan nilai rata-rata akurasi sebesar 96,25%. Sehingga dapat disimpulkan bahwa klasifikasi yang dihasilkan sudah baik. Kelebihan dari metode *Naïve Bayes Classifier* ini adalah algoritma ini memiliki akurasi yang tinggi [7].

IV. KESIMPULAN

Naïve Bayes Classifier dapat digunakan untuk klasifikasi teks berita berbahasa Indonesia melalui beberapa tahapan yaitu proses memasukkan data yang berupa teks dengan format .txt, tahap praproses data (*case folding*, *tokenization*, *stopword removal*, dan *stemming*), tahap pembobotan kata dengan *term frequency*, dan tahap klasifikasi dengan metode *Naïve Bayes Classifier*.

Tingkat akurasi optimal yang dihasilkan adalah 100% pada klasifikasi kategori dalam pengujian pertama dan kedua. Pengujian pertama dilakukan dengan data latih sejumlah 180 dan data uji 20. Pada pengujian kedua dilakukan dengan data latih sejumlah 160 dan data uji 40. Dari kedua pengujian tersebut semua data uji berhasil diklasifikasikan dengan benar.

DAFTAR PUSTAKA

[1] A. N. Chy, H. Seddiqui, and S. Das, *Bangla news classification using naive Bayes classifier*. 2014.

[2] I. G. A. Socrates, A. L. Akbar, and M. S. Akbar, "Optimasi Naïve Bayes Dengan Pemilihan Fitur Dan Pembobotan Gain Ratio," vol. 7, no. 1, pp. 22–30, 2016.

[3] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Mach. Learn.*, vol. 29, no. 2–3, pp. 103–130, 1997.

[4] J. Aldrich, "R . A . Fisher on Bayes and Bayes ' Theorem," no. 1, pp. 161–170, 2008.

[5] A. Nafalski and A. P. Wibawa, "Machine translation with Javanese speech levels' classification," *Inform. Autom. Pomiary w Gospod. i Ochr. Środowiska*, 2016.

[6] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *2008 IEEE/ACS international conference on computer systems and applications*, 2008, pp. 108–115.

[7] I. Rish, "An empirical study of the naive Bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 2001, vol. 3, no. 22, pp. 41–46.