Pasinformatik, Volume 5, Issue 2, July 2026
E-ISSN : 2986-5360

Klasifikasi Otomatis Teks Berita Berbahasa Indonesia Menggunakan Algoritma *Naïve Bayes Classifier*

Agus Hexagraha[*], Asep Dede Ruswanda[**]

Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pasundan
Jln. Dr. Setiabudhi no. 193 Bandung, Jawa Barat
[*]hexagraha@unpas.ac.id, [**]asepdedekoswara@gmail.com

Abstrak. Teks berbahasa alami yang tidak terstruktur mendominasi sebagian besar informasi digital saat ini, sehingga muncul kebutuhan mendesak untuk melakukan ekstraksi dan kategorisasi informasi secara otomatis. Penelitian ini menerapkan algoritma Naïve Bayes Classifier untuk mengklasifikasikan dokumen teks berita ke dalam tiga kategori utama, yaitu Ekonomi, Olahraga, dan Pendidikan. Proses klasifikasi diawali dengan tahapan preprocessing dokumen yang meliputi case folding, tokenizing, penghilangan stopwords, dan stemming. Dataset penelitian mencakup total 150 data latih (50 dokumen per kategori) dan dievaluasi menggunakan 30 dokumen uji yang didistribusikan secara merata. Hasil pengujian menunjukkan bahwa algoritma Naïve Bayes Classifier mampu menghasilkan tingkat akurasi yang tinggi, yaitu sebesar 93.33% dengan tingkat kesalahan (error rate) sebesar 6.67%. Evaluasi mendalam menggunakan Confusion Matrix menunjukkan nilai Precision, Recall, dan F1-Score rata-rata berada pada angka 93.33%, membuktikan bahwa metode ini sangat efektif dan efisien untuk menangani klasifikasi teks berskala besar.

Kata Kunci: *Text Mining, Klasifikasi Teks, Naïve Bayes Classifier, Confusion Matrix, Berita Berbahasa Indonesia.*

I. PENDAHULUAN

Perkembangan teknologi informasi yang masif memicu terjadinya fenomena *information overload*, di mana volume data tekstual meningkat tajam secara eksponensial sementara kemampuan manusia untuk memproses informasi cenderung konstan [1]. Diperkirakan sekitar 80% data online tersedia dalam bentuk format teks tidak terstruktur dan mayoritas tidak dilengkapi dengan metadata yang memadai [3]. Oleh karena itu, *text mining* atau *knowledge discovery in text* (KDT) hadir sebagai solusi untuk menganalisis teks berbahasa alami guna menemukan pola atau informasi baru yang sebelumnya tidak diketahui [1], [5]. Salah satu operasi fundamental dalam *text mining* adalah kategorisasi teks (*supervised learning*), yaitu aktivitas memberikan label kategori subjek pada dokumen baru berdasarkan karakteristik yang diusulkan oleh data latih berlabel [1]. Implementasi nyata dari kategorisasi teks ini adalah sistem klasifikasi berita otomatis. Di antara berbagai rumpun algoritma *text classifier* kuantitatif, *Naïve Bayes Classifier* (NBC) merupakan metode klasifikasi statistik yang sangat populer. Algoritma ini dipilih karena memiliki beberapa kelebihan utama, seperti proses komputasi yang cepat, struktur algoritma yang sederhana, kokoh terhadap *noise*, serta memiliki tingkat akurasi yang tinggi saat diimplementasikan pada data berukuran besar. Penelitian ini bertujuan menerapkan NBC untuk mengelompokkan teks berita ke dalam domain Ekonomi, Olahraga, dan Pendidikan, sekaligus menyajikan analisis performa metrik evaluasi secara komprehensif.

II. METODE PENELITIAN

Penelitian ini menerapkan kerangka kerja Knowledge Discovery in Database (KDD) yang dikhususkan untuk pemrosesan data teks (Text Mining). Tahapan penelitian disusun secara sistematis untuk mengubah data mentah berupa artikel berita bahasa Indonesia menjadi informasi terstruktur yang siap diklasifikasikan menggunakan algoritma Naive Bayes Classifier. Alur metodologi dalam penelitian ini dibagi menjadi tiga tahapan utama: pra-

pemrosesan teks (text preprocessing), pemodelan klasifikasi, dan evaluasi kinerja. Ketiga tahapan tersebut dibantu dengan cara membangun aplikasi untuk melakukan pemrosesan tersebut. Di bagian akhir dari penelitian ini disampaikan kesimpulan yang dihasilkan termasuk temuan-temuannya.

2.1 Preprocessing Dokumen

Tahap ini bertujuan untuk melakukan analisis sintaktik, mereduksi dimensi kata, dan mengubah dokumen teks bebas menjadi bentuk representasi data terstruktur (*Bag-of-Words*) [3]. Langkah-langkah pengolahan awal ini meliputi:

a. Case Folding: Mengubah seluruh huruf di dalam dokumen berita menjadi huruf kecil (*lowercase*) secara seragam.
b. Tokenizing: Memotong rangkaian kalimat dalam teks berita menjadi satuan kata tunggal atau token.
c. Stopwords Removal (Filtering): Membuang kata-kata umum (seperti "yang", "di", "dan", "dari") yang sering muncul namun tidak memiliki bobot informatif dalam penentuan kategori.
d. Stemming: Mengubah setiap kata berimbuhan menjadi kata dasarnya untuk menghilangkan variasi morfologi kata penentu.

2.2 Pemodelan Naïve Bayes Classifier

Algoritma *Naïve Bayes* beroperasi menggunakan Teorema Bayes dengan asumsi independensi yang kuat bahwa efek suatu nilai atribut/kata pada kelas yang diberikan adalah bebas dari atribut lainnya [3]. Hipotesis probabilitas maksimum dari kategori target (v_{MAP}) ditentukan melalui formula berikut:

$$v_{MAP} = \arg\max_{vj \in V} P(v_j) \prod_i P(a_i \mid v_j)$$

Dimana $P(v_j)$ mewakili probabilitas awal dari kategori berita v_j, sedangkan $P(w_i \mid v_j)$ melambangkan probabilitas kemunculan kata conditional w_i pada kategori v_j. Perhitungan probabilitas kata memanfaatkan estimasi *Laplace Smoothing* guna mengantisipasi kemunculan nilai probabilitas nol saat ada kata baru pada data uji yang tidak terekam di data latih [2].

Dimana v_{MAP} adalah nilai keluaran hasil klasifikasi *naive bayes.*

$p(v_j)$ dan probabilitas kata w_k untuk setiap kategori $p(w_k \mid v_j)$ dihitung pada saat pelatihan.

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

di mana :

$|docs_j|$ adalah jumlah kata pada kategori $_j$.

$|contoh||$ adalah jumlah kata yang digunakan dalam pelatihan.

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

di mana:

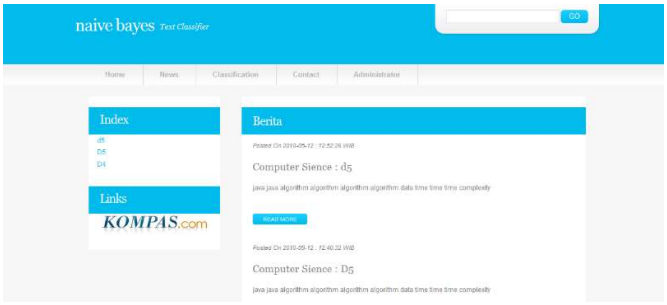
n_k adalah jumlah kemunculan kata w_k pada kategori v_j .

n adalah jumlah semua kata pada kategori v_j .

$|kosakata|$ adalah jumlah kata yang unik (*distinc*) pada semua data latih.

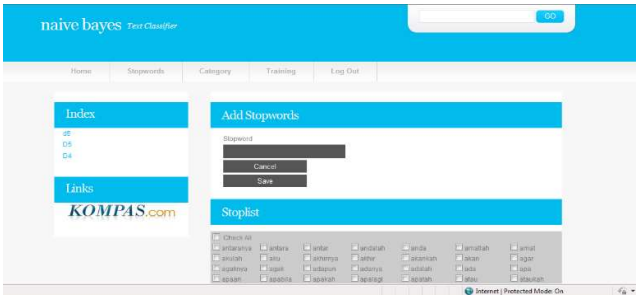
III. HASIL DAN PEMBAHASAN

Dataset penelitian yang digunakan mencakup total 150 data latih (50 dokumen per kategori) dan dievaluasi menggunakan 30 dokumen uji yang didistribusikan secara merata. Dalam algoritma Naïve Bayes Classifier, "pola" yang dihasilkan dari proses pembelajaran (training) tidak berupa aturan logika (if-then) atau pohon keputusan, melainkan berupa Basis Pengetahuan Probabilitas (Model Probabilitas). Gambar 1 sampai dengan 5 merupakan tampilan dari aplikasi dari klasifikasi berita.



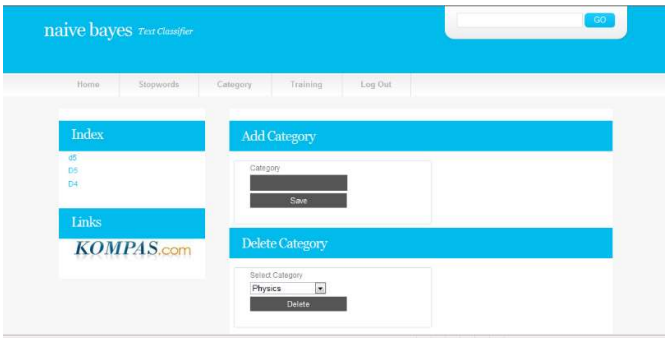
Gambar 1. Tampilan halaman lihat berita

Gambar 1 menampilkan Halaman Lihat Berita: Menampilkan daftar dokumen berita yang tersimpan di dalam basis data beserta detail isi teks dan kategorinya. Halaman ini berfungsi untuk pengelolaan data berita (Create, Read, Update, Delete) baik untuk kebutuhan data latih maupun data uji.



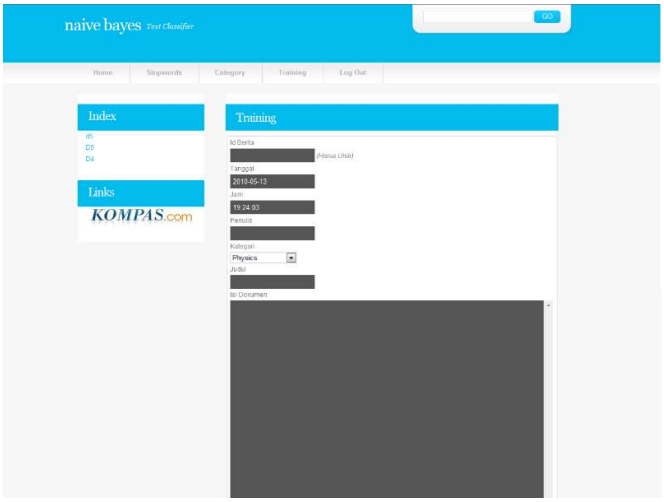
Gambar 2. Tampilan halaman *stopwords*

Gambar 2. menampilkan Halaman Stopwords: Menampilkan daftar kata buang (stopwords) bahasa Indonesia. Halaman ini digunakan untuk mengelola kata-kata umum (seperti "yang", "di", "dan") yang akan dieliminasi pada tahap preprocessing agar tidak memengaruhi bobot probabilitas kata.



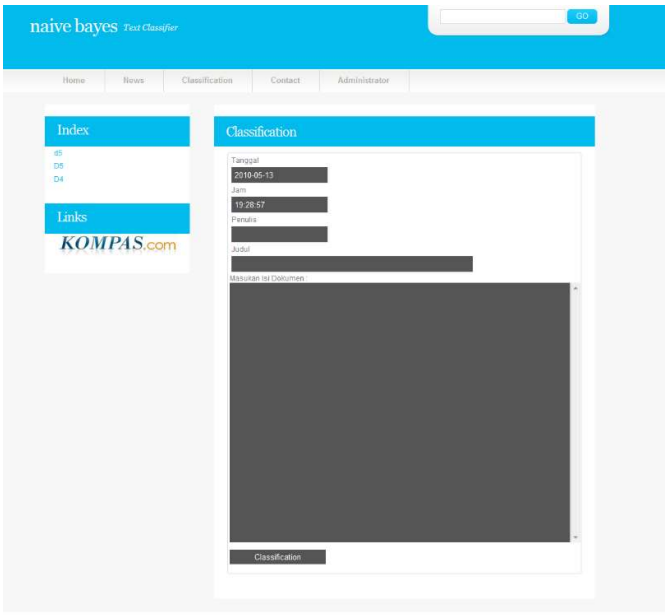
Gambar 3. Tampilan halaman kategori

Gambar 3 menampilkan Halaman Kategori: Menampilkan daftar kelas atau label target klasifikasi (Ekonomi, Olahraga, dan Pendidikan). Halaman ini berfungsi sebagai referensi master data kategori yang digunakan oleh algoritma.



Gambar 4. Tampilan halaman *training*

Gambar 4. menampilkan Halaman Training: Menjalankan proses pembelajaran (training) dari 150 data latih. Halaman ini memproses tahapan preprocessing hingga menghitung probabilitas prior dan likelihood kata untuk membentuk tabel basis pengetahuan (knowledge base).



Gambar 5. Tampilan halaman *classification*

Gambar 5 menmpilkan Halaman Classification: Antarmuka utama untuk menguji atau mengklasifikasikan dokumen berita baru. Pengguna memasukkan teks berita, kemudian sistem menghitung nilai probabilitas dan menampilkan hasil prediksi kategorinya secara otomatis.

Pengujian performa sistem dilakukan menggunakan 30 dokumen uji (*data test*) eksternal yang didistribusikan secara berimbang, yakni masing-masing 10 dokumen untuk kelas Ekonomi, Olahraga, dan Pendidikan. Dari total 30 dokumen uji yang dimasukkan ke dalam mesin pengklasifikasi, sebanyak 28 dokumen berhasil diprediksi secara tepat sesuai dengan label aktualnya, sedangkan 2 dokumen lainnya mengalami salah klasifikasi. Perhitungan performa makro sistem menghasilkan parameter sebagai berikut:

a. **Tingkat Akurasi (*Accuracy*):**

$$Accuracy = \frac{numberOfCorrect\,\mathrm{Pr}ediction}{totalNumberOf\,\mathrm{Pr}ediction} = (28/30) * 100\% = 93.33\%$$

b. **Tingkat Kesalahan (*Error Rate*):**

$$ErrorRate = \frac{numberOfWrong\,\mathrm{Pr}ediction}{totalNumberOf\,\mathrm{Pr}ediction} = (2/30) * 100\% = 6.67\%$$

Untuk mengukur performa penyebaran prediksi pada setiap kelas target, dibentuk pemetaan evaluasi menggunakan tabel *Confusion Matrix* di bawah ini.

Tabel 1. Confusion Matrix

Kategori Aktual \ Prediksi	Ekonomi	Olahraga	Pendidikan	Total Aktual
Ekonomi	9	0	1	10
Olahraga	0	10	0	10
Pendidikan	1	0	9	10
Total Prediksi	10	10	10	30

Berdasarkan data matriks di atas, kalkulasi metrik efektivitas klasifikasi tingkat kategori (*Precision, Recall,* dan *F1-Score*) dirangkum dalam tabel berikut.

Tabel 2. Metrik Efektifitas Klasifikasi Tingkat Kategori

Kategori Berita	Precision	Recall	F1-Score
Ekonomi	90%	90%	90%
Olahraga	100%	100%	100%
Pendidikan	90%	90%	90%
Rata-rata (Average)	93.33%	93.33%	93.33%

Analisis Kasus Salah Klasifikasi

Penelitian ini melakukan analisis mendalam terhadap 2 dokumen uji yang gagal dikategorikan secara tepat oleh sistem untuk mengetahui batasan performa algoritma:

1. Kasus Dokumen Ekonomi Diprediksi Sebagai Pendidikan
   Dokumen ini secara aktual membahas topik Anggaran Pendapatan dan Belanja Negara (APBN) untuk pendanaan beasiswa perguruan tinggi. Secara esensi substansi, berita ini masuk ke ranah Ekonomi. Namun, karena *Naïve Bayes* menghitung probabilitas kata secara independen tanpa memedulikan konteks frasa [3], kemunculan kata berciri khas pendidikan dengan frekuensi tinggi (seperti *"mahasiswa"*, *"beasiswa"*, dan *"kampus"*) menghasilkan nilai akumulasi *likelihood* posterior kelas Pendidikan yang lebih dominan, sehingga mengungguli bobot kata bertema Ekonomi.
2. Kasus Dokumen Pendidikan Diprediksi Sebagai Ekonomi
   Artikel berita ini menelaah isu komersialisasi pendidikan terkait kenaikan tarif Uang Kuliah Tunggal (UKT). Kekeliruan prediksi terjadi akibat jurnalis banyak menggunakan terminologi bernuansa finansial secara berulang,

seperti token *"tarif"*, *"biaya"*, *"inflasi"*, dan *"daya beli"*. Kata-kata tersebut memiliki bobot probabilitas awal yang sangat masif pada basis data latih kategori Ekonomi, memicu sistem condong memilih kelas Ekonomi sebagai hasil akhir keputusan [3].

Hasil analisis ini membuktikan karakteristik dasar *Naïve Bayes* yang sangat rentan dan sensitif terhadap *overlapping vocabulary* (irisan kosakata) antar-domain ketika sebuah dokumen ditulis menggunakan kombinasi gaya bahasa lintas disiplin ilmu [2].

IV. KESIMPULAN

Penelitian ini berhasil membuktikan bahwa algoritma *Naïve Bayes Classifier* memiliki efektivitas yang sangat tinggi untuk melakukan kategorisasi teks berita berbahasa Indonesia secara otomatis, ditunjukkan dengan capaian nilai akurasi makro sebesar 93.33%. Saluran kategori Olahraga mencatatkan performa paling sempurna (100.00%) karena memiliki karakteristik kosakata yang unik dan terisolasi. Kendala utama algoritma ini terletak pada sensitivitasnya terhadap dokumen yang memuat *overlapping vocabulary* lintas domain. Untuk penelitian selanjutnya, disarankan untuk mengintegrasikan tahapan seleksi fitur (*Feature Selection*) seperti *Information Gain* atau *Chi-Square* sebelum pemodelan guna mereduksi *noise* dari kata-kata yang bermakna ganda.

V. REFERENSI

[1] Defiyanti, S., & Jajuli, M. (2015). Integrasi Metode Klasifikasi Dan Clustering dalam Data Mining. *Konferensi Nasional Informatika (KNIF) 2015*, 39-44.
[2] Muhamad, H., Prasojo, C. A., Sugianto, N. A., Surtiningsih, L., & Cholissodin, I. (2017). Optimasi Naïve Bayes Classifier Dengan Menggunakan Particle Swarm Optimization Pada Data Iris. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, 4(3), 180-184.
[3] Kusrini , Taufiq Luthfi, Emha. (2009). ”Algoritma Data Mining”, Edisi I, CV. Andi Offset, Yogyakarta,
[4] Santosa, Budi.(2007). Teknik Pemanfaatan Data Untuk Keperluan Bisnis, edisi I, Yogyakarta, Graha Ilmu,
[5] Tan, Pang-ning., Steinbach, Michael., dan Kumar,Vipin. (2009). Introduction To Data Mining. Addison-Wesley