Integrated Journal of Applied Technology and Informatics
Journal homepage: https://journal2.uinssc.ac.id/index.php/IJATI

Evaluation of Classical Machine Learning Methods and TextCNN in Indonesian News Classification

Evaluasi Metode Machine Learning Klasik dan TextCNN dalam Klasifikasi Berita Berbahasa Indonesia

Muhammad Rafly Saputra, Dika Hasan Nugroho, Muhammad Syifaulqulub Hakim, Gina Khayatun Nufus
Program Studi Informatika, Universitas Islam Negeri Siber Syekh Nurjati Cirebon, Cirebon, 45132, Indonesia
raflybiasa20@gmail.com; dikan0914@gmail.com; m.syifaulqulubhakim@gmail.com; *ginakhn@uinssc.ac.id;*

ARTICLE INFO
Article history:
Received 2026-06-30
Revised 2026-06-30
Accepted 2026-06-30

CORRESPONDING AUTHOR
Gina Khayatun Nufus
ginakhn@uinssc.ac.id

ABSTRACT
The massive growth of digital news demands reliable automated classification systems. However, systematic evaluations of optimal methods for the Indonesian language, particularly under data-constrained environments, remain scarce. This study evaluates and compares conventional Machine Learning and Deep Learning approaches in classifying multiclass Indonesian news articles. Utilizing a dataset of 1,199 articles scraped from Detik.com, this research applies an automated labeling scheme based on URL subdomain structures. Three methods are comparatively tested: TF-IDF-based Multinomial Naïve Bayes (MNB) and Logistic Regression (LR), against a Deep Learning TextCNN architecture, using a combined title and three-paragraph feature representation. Experimental results demonstrate that Logistic Regression achieves the highest performance (86.67% accuracy, 78.84% F1-Macro), significantly outperforming TextCNN (67.08% accuracy) due to data scarcity constraints. Error analysis reveals that model misclassifications are concentrated within semantically close categories, such as misclassifying Finance as News due to lexical overlap in government policy discussions. These findings underscore that for small-scale Indonesian corpora, classical TF-IDF methods offer superior classification efficacy and computational efficiency compared to complex Deep Learning architectures.
***Keyword**: News Classification, Indonesian Language, Logistic Regression, TextCNN, Small-scale Corpus*

ABSTRAK
Pertumbuhan masif berita digital menuntut sistem klasifikasi otomatis yang andal. Namun, evaluasi sistematis mengenai metode optimal dalam konteks bahasa Indonesia—khususnya pada kondisi keterbatasan data—masih minim dilakukan. Penelitian ini membandingkan kinerja metode *Machine Learning* konvensional dan *Deep Learning* dalam mengklasifikasikan kategori berita berbahasa Indonesia. *Dataset* sebanyak 1.199 artikel dihimpun dari Detik.com melalui *web scraping* dengan skema pelabelan otomatis berbasis struktur *subdomain* URL. Tiga metode diuji secara komparatif: *Multinomial Naïve Bayes* (MNB) dan *Logistic Regression* (LR) berbasis TF-IDF, serta arsitektur *Deep Learning TextCNN*, menggunakan fitur gabungan judul dan tiga paragraf pertama. Hasil eksperimen menunjukkan bahwa *Logistic Regression* mencapai performa tertinggi dengan akurasi 86,67% dan *F1-Macro* 78,84%, mengungguli *Text*CNN secara signifikan (akurasi 67,08%) yang terkendala keterbatasan volume data latih. Analisis kesalahan mengungkap bahwa klasifikasi model terkonsentrasi pada kategori dengan kedekatan semantik tinggi, seperti transisi *Finance* menjadi *News* akibat tumpang tindih diksi kebijakan pemerintah. Temuan ini menegaskan bahwa untuk korpus bahasa Indonesia berskala kecil, metode klasik berbasis TF-IDF menunjukkan efektivitas dan efisiensi komputasi yang lebih unggul dibandingkan pendekatan *Deep Learning* yang kompleks.
Kata Kunci: Klasifikasi Berita, Bahasa Indonesia, TF-IDF, *Logistic Regression*, *Text*CNN, *Natural Language Processing*

1. PENDAHULUAN

Perkembangan tekno*log*i informasi dalam era digital saat ini telah merubah fundamental cara masyarakat dalam memproduksi, menyebarkan, dan mengonsumsi informasi. Portal berita *online* terkemuka seperti detik.com menghasilkan ratusan hingga ribuan artikel setiap harinya yang mencakup beragam ranah kehidupan, mulai dari berita nasional, ekonomi, olahraga, tekno*log*i, kesehatan, hingga otomotif. Pertumbuhan data teks yang sangat masif ini melahirkan tantangan baru berupa bagaimana mengorganisasi dan mengelompokkan informasi secara cepat, akurat, dan efisien. Proses pengelompokan artikel ke dalam kategori yang sesuai secara manual menggunakan tenaga manusia memiliki keterbatasan yang sangat signifikan dari segi waktu, biaya, dan konsistensi. Faktor kejenuhan manusia dapat memicu terjadinya *human error* berupa salah pengkategorian dokumen teks. Oleh karena itu, diperlukan suatu mekanisme otomatisasi yang memanfaatkan perkembangan di bidang *Natural Language Processing* (NLP) dan *Machine Learning* untuk mampu mengklasifikasikan topik berita secara presisi berdasarkan karakteristik linguistik teks[1]. Studi mengenai klasifikasi teks bahasa Indonesia telah banyak dilakukan oleh peneliti terdahulu menggunakan berbagai pendekatan. Beberapa penelitian berfokus pada efisiensi ekstraksi fitur berbasis pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF) yang dikombinasikan dengan pengklasifikasi probabilistik konvensional seperti *Naïve Bayes* maupun algoritma berbasis jarak seperti *K-Nearest Neighbors* (K-NN)[2]. Penelitian terbukti menunjukkan hasil bervariasi bergantung pada karakteristik *dataset* dan kedalaman tahapan *preprocessing* yang diimplementasikan[3]. Meskipun algoritma klasik ini terbukti efisien pada *dataset* berukuran kecil hingga menengah, munculnya arsitektur *Deep Learning* seperti *Convolutional Neural Network* untuk klasifikasi teks (*Text*CNN) memberikan perspektif baru[4]. *Text*CNN mampu menangkap pola spasial dan hubungan n-gram lokal dalam teks melalui operasi konvolusi dengan kernel yang bervariasi. Namun, efektivitas penggunaan arsitektur kompleks ini pada repositori data berita berbahasa Indonesia yang berskala menengah masih membutuhkan eksplorasi dan perbandingan yang komprehensif guna memahami *trade-off* antara performa komputasi dan akurasi klasifikasi[5].

2. METODE

Sistem klasifikasi otomatis yang dirancang dalam penelitian ini secara terstruktur dibangun atas beberapa tahapan sekuensial yang meliputi:

a. Pengumpulan data mentah melalui *web scraping*.
b. Pelabelan otomatis dan penapisan kualitas *dataset*.
c. Pra-pemrosesan teks (*text preprocessing*).
d. Ekstraksi fitur teks berbasis TF-IDF dan representasi *token* numerik.
e. Pelatihan model *multi*-algoritma.
f. Evaluasi metrik performa klasifikasi.

2.1. Teknik Pengumpulan Data

Proses pengumpulan data dilakukan menggunakan teknik penambangan *web* (*web scraping*) pada halaman indeks portal berita Detik.com. Implementasi skrip *scraping* dikembangkan menggunakan bahasa pemrograman Python dengan memanfaatkan pustaka *Requests* untuk menangani protokol HTTP *request* dan pustaka *BeautifulSoup* untuk melakukan *parsing* dokumen HTML[6]. Komponen data yang berhasil diekstrak dari setiap artikel meliputi: judul artikel, tiga paragraf pertama, keseluruhan isi teks secara penuh (*full text*), URL artikel, serta tanggal publikasi. Strategi pelabelan kelas target (*ground truth*) pada *dataset* ini mengadopsi skema otomatis yang mengeksploitasi struktur penamaan *subdomain* pada URL berita. Setiap artikel yang diekstrak dipetakan secara otomatis ke dalam kategori yang bersesuaian berdasarkan pola URL, sebagai contoh: artikel dengan tautan 'https://news.detik.com/...' secara otomatis dikelompokkan ke dalam kategori *'News'*, sementara tautan 'https://finance.detik.com/...' dimasukkan ke dalam kategori *'Finance'*.

2.2. Karakteristik dan Distribusi Dataset

Proses pengumpulan awal berhasil menjaring sebanyak 1.200 artikel mentah. Setelah dilakukan tahapan *data cleaning*, terdeteksi terdapat 1 rekaman data duplikat dan 2 rekaman yang kehilangan nilai pada atribut tanggal (*missing values*), sehingga total *dataset* bersih yang digunakan dalam eksperimen ini berjumlah 1.199 artikel. Distribusi sebaran data pada masing-masing kategori dapat dilihat secara terperinci pada Tabel 1. Kemunculan kategori '20' diidentifikasi sebagai sebuah

anomali atau *noise* yang bersumber dari kesalahan *parsing* ekspresi reguler saat proses ekstraksi URL teks dilakukan. Kategori minoritas ini sengaja dipertahankan di dalam repositori *dataset* untuk menguji tingkat ketahanan (*robustness*) dari masing-masing algoritma klasifikasi dalam menangani kondisi data yang tidak seimbang (*imbalanced dataset*).

Tabel 1. Sebaran Distribusi Kategori pada *Dataset* Akhir.

Kategori	Jumlah
News	200
Health	200
Inet	200
Oto	199
Finance	185
Sport	185
20 (anomali)	30

2.3. Tahapan Pra-Pemrosesan Teks (Text Preprocessing)

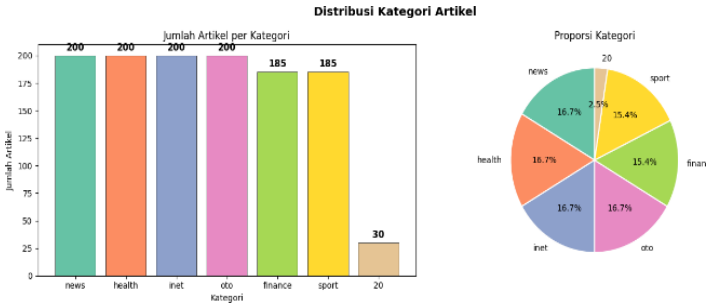
Data teks berita hasil *scraping* bersifat tidak terstruktur dan mengandung banyak *noise linguistik* yang dapat mereduksi tingkat akurasi model[7]. Oleh karena itu, diimplementasikan rangkaian tahapan *text preprocessing* baku yang meliputi:

a. *Lowercase Conversion*: Mengubah seluruh huruf kapital dalam teks menjadi huruf kecil guna menghindari redundansi *token*.
b. Penghapusan URL: Menghilangkan *string* tautan *website* yang tidak berkorelasi dengan topik utama berita.
c. Penghapusan Angka dan Tanda Baca: Mengeliminasi karakter numerik serta simbol-simbol punktuasi seperti tanda koma, titik, dan tanda kurung;
d. Penghapusan Karakter Khusus: Membersihkan simbol non-ASCII dan spasi berlebih (*whitespace*) yang tidak diperlukan. Sebagai ilustrasi konkret, transformasi teks asli berjudul 'AHY Kejar Proyek *Giant Sea Wall* Pantura: Semoga Tahun Depan Konsep Matang' setelah melewati jalur *preprocessing* berubah menjadi deretan *token* bersih: 'ahy kejar proyek *giant sea wall* pantura semoga konsep matang'.

2.4. Eksplanatory Data Analysis (EDA)

Adapun *eksplanatory* data analisis yang dilakukan yaitu distribusi kategori, panjang dokumen serta kosakata khas dan *WordCloud,* sebagai berikut:

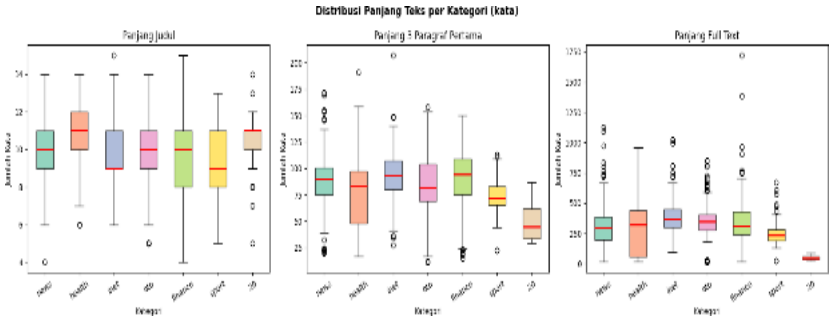
a. Distribusi kategori.
Hasil visualisasi menunjukkan distribusi kategori relatif seimbang pada enam kategori utama dengan jumlah sekitar 185–200 artikel per kategori. Namun ditemukan kategori minoritas "20" yang hanya memiliki 30 artikel sehingga berpotensi memengaruhi performa model.



Gambar 1. Distribusi Kategori.

b. Panjang dokumen
Riset ini melakukan komparasi panjang teks berdasarkan beberapa variasi masukan dokumen. Hasil kalkulasi statistik deskriptif menunjukkan bahwa komponen 'Judul' memiliki panjang rata-rata sebanyak 7 kata per artikel. Skema gabungan 'Judul + 3 Paragraf pertama' memiliki panjang rata-rata 57 kata, komponen '3 Paragraf Pertama' saja memiliki

rata-rata 50 kata, sedangkan dokumen penuh ('*Full Text*') memiliki panjang rata-rata mencapai 181 kata per artikel. Pemetaan statistik panjang teks penuh untuk masing-masing kategori tersaji secara rinci pada Gambar 2.



Gambar 2. Boxplot Panjang Artikel

c. Kosakata khas dan *WordCloud*

Melalui pemodelan kata dominan (*Top Words*), ditemukan indikasi bahwa setiap kategori memiliki sebaran leksikal unik yang merepresentasikan domain topiknya masing-masing. Kategori *'Sport'* didominasi oleh kemunculan *token* khas seperti 'motogp', 'balapan', 'marquez', dan 'juara'. Kategori *'Health'* memiliki kecenderungan pemakaian kata 'kesehatan', 'pasien', 'penyakit', dan 'obat'. Kategori *'Finance'* didominasi oleh kata 'ekonomi', 'bank', 'harga', dan 'keuangan', sedangkan kategori 'Oto' dicirikan oleh *token* 'motor', 'mobil', 'kendaraan', dan 'pasar'. Perbedaan profil leksikal yang kontras ini mengonfirmasi secara empiris bahwa *dataset* memiliki tingkat diskriminasi fitur yang memadai untuk diproses oleh model klasifikasi.

Gambar 3. Kosakata Khas dan *WordCloud*

2.5. Eksperimen Fitur

Guna mengidentifikasi konfigurasi data masukan yang paling efisien, dilakukan eksperimen pengujian terhadap empat variasi panjang fitur teks masukan menggunakan ekstraksi fitur TF-IDF *Vectorizer* yang diujikan secara konsisten menggunakan *baseline* model *Multinomial Naïve Bayes*. Perbandingan nilai *Accuracy* dan F1-*Macro* dari eksperimen variasi fitur ini dirangkum dalam Tabel 2.

Tabel 2. Hasil Eksperimen Pengaruh Variasi Fitur Teks Terhadap Performa Model

Variasi Fitur	***Accuracy***	**F1 *Macro***
Judul Saja	80.00%	75.95%
Judul + 3 Paragraf	84.17%	72.90%
3 Paragraf Saja	83.75%	72.50%
Full Text	84.58%	73.39%

Hasil eksperimen menunjukkan bahwa pemanfaatan komponen '*Full Text*' memang menghasilkan akurasi tertinggi sebesar 84,58%. Namun, selisih performa akurasi dengan variasi 'Judul+ 3 Paragraf Pertama' sangatlah tipis, yaitu hanya sebesar 0,41%. Mempertimbangkan aspek efisiensi komputasi, ukuran dimensi matriks *sparse* yang lebih ringkas, serta kecepatan waktu inferensi, maka komponen 'Judul + 3 Paragraf Pertama' dipilih sebagai representasi fitur teks utama untuk diimplementasikan pada tahap pelatihan dan komparasi model lanjutan.

2.6. Impementasi Model

Dalam implementasi model menggunakan metode *machine learning* (*multinomial naive bayes* dan *logistic regression*) serta metode *text*CNN untuk arsitektur *deep learning*. Adapun masing-masing penjelasan dari metode yang digunakan yaitu:

a. *Multinomial Naive Bayes*
   Model pertama dibangun berbasiskan *Multinomial Naïve Bayes* yang dipadukan dengan pembobotan TF-IDF *Vectorizer*. Model ini bekerja menggunakan teori probabilitas dengan menghitung frekuensi kemunculan *token* dalam suatu kelas dokumen, di bawah asumsi independensi fitur yang kuat (kondisi *conditional independence*). Model ini sangat efisien dari sudut pandang waktu eksekusi pelatihan[8].
b. *Logistic Regression*
   Model kedua menggunakan algoritma *Logistic Regression* yang dikonfigurasikan untuk kasus *multi*kelas (*multinomial*) dengan fungsi aktivasi *softmax* untuk menghasilkan distribusi nilai probabilitas estimasi kelas[9]. Pendekatan ini memetakan hubungan linear antara fitur-fitur berbobot TF-IDF dengan *log-odds* dari masing-masing kategori berita.
c. textCNN

Model ketiga mengadopsi pendekatan *Deep Learning* melalui arsitektur TextCNN. Dokumen teks terlebih dahulu di transformasikan ke dalam bentuk *token* urutan numerik (*sequence tokenization*). Komponen penyusun arsitektur TextCNN dalam penelitian ini dapat di lihat pada Tabel 3.

Tabel 3. TextCNN Arsitektur

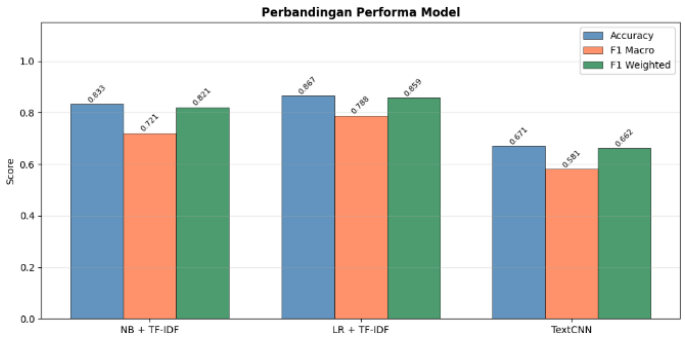
Layer	Penjelasan
Embedding Layer	Dimensi representasi vektor sebesar 128 untuk ukuran kosakata (*Vocabulary Size*) sebanyak 11.275 *token*.
Convolutional 1D	Tiga buah *Convolutional* 1D Layer paralel yang masing-masing menerapkan ukuran filter (kernel *size*) bervariasi yaitu 2, 3, dan 4 untuk menangani ekstraksi pola variasi n-gram teks.
Max Pooling	*Global Max Pooling Layer* untuk mereduksi dimensi dan mengambil fitur spasial paling dominan.
Fully Connected (Dense) Layer	regulasi *Dropout* sebesar 0,4 untuk mencegah *overfitting*.
Output Layer	*Output Layer* berbasis *Softmax*. Total parameter yang dilatih pada model ini berjumlah 1.692.167 parameter dengan durasi pelatihan selama 15 *epoch.*

3. HASIL DAN PEMBAHASAN

Setelah proses pelatihan selesai dilaksanakan terhadap komponen fitur 'Judul + 3 Paragraf Pertama', dilakukan pengujian menggunakan data uji independen guna memperoleh gambaran objektif mengenai kemampuan generalisasi dari masing-masing model.

3.1. Komparasi Kinerja Antar Model

Metrik evaluasi yang digunakan sebagai indikator performa meliputi *Accuracy*, F1-*Macro* (untuk mengukur performa rata-rata tanpa terpengaruh ketidakseimbangan data), serta F1-Weighted. Ringkasan hasil evaluasi performa akhir disajikan dalam Gambar 4. Berdasarkan data komparatif pada Gambar 4, model *Logistic Regression* (LR) menempati urutan pertama sebagai model dengan performa tertinggi, mencatatkan nilai akurasi sebesar 86,67% dan nilai F1-*Macro* 78,84%.



Gambar 4. Komparasi Kinerja Metrik Evaluasi Akhir Antar Model

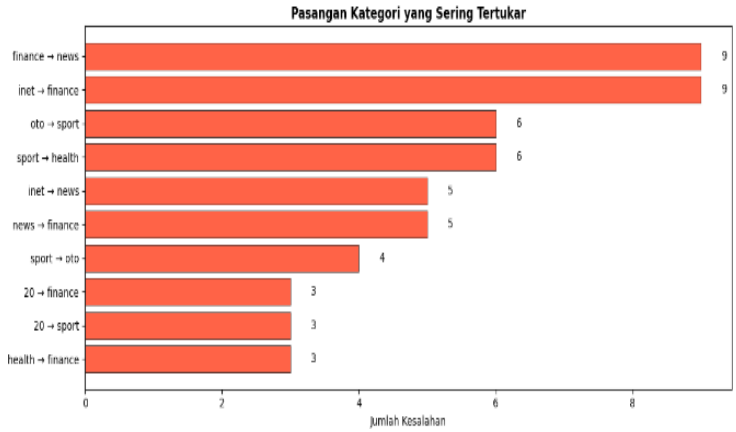
Model *Multinomial Naïve Bayes* juga memperlihatkan performa yang solid dan bersaing ketat dengan raihan akurasi sebesar 83,33%. Di sisi lain, model *Deep Learning Text*CNN menunjukkan performa yang kurang optimal dengan nilai akurasi terendah di angka 67,08%. Rendahnya capaian akurasi *Text*CNN ini diasumsikan terjadi karena *volume dataset* bahasa Indonesia yang digunakan tergolong berskala kecil (*small-to-medium dataset*, yaitu 1.199 artikel). Arsitektur *deep learning* yang memiliki jutaan parameter cenderung membutuhkan kuantitas *supply* data latih yang jauh lebih masif agar terhindar dari fenomena *overfitting* dan mampu membentuk representasi *embedding* kata yang matang.

3.2. Analisis Matriks Kebingungan (Confusion Matrix)

Analisis mendalam menggunakan *confusion matrix* pada model berbasis TF-IDF menunjukkan bahwa sistem mampu mengenali kelas dengan sangat baik pada beberapa kategori spesifik. Tingkat pengenalan akurasi lokal tertinggi dicapai pada kategori *Health* (95,0%), kategori Inet (97,5%), serta kategori *Sport* (97,3%). Keberhasilan ini didorong oleh karakteristik leksikal unik dan spesifik dari ketiga domain topik tersebut yang jarang tumpang tindih dengan domain lainnya. Sebaliknya, kategori *'News'* teridentifikasi memiliki tingkat kerentanan misklasifikasi yang paling tinggi (kesalahan lokal mencapai 35,0%). Hal ini disebabkan oleh sifat dari kategori *'News'* itu sendiri yang mencakup informasi umum (*general news*), sehingga kosa kata yang terkandung di dalamnya bersifat heterogen dan sering kali memuat entitas atau diksi yang beririsan langsung dengan rumpun kategori ekonomi, tekno*log*i, maupun olahraga.

3.3. Analisis Pola Kesalahan Klasifikasi (Error Analysis)

Penelitian ini berhasil mengidentifikasi adanya tiga pasangan kategori utama (top-3 confused pairs) yang paling sering mengalami salah klasifikasi secara konsisten oleh model[10]. Pasangan kategori beserta frekuensi kesalahan tersebut dirangkum dalam Gambar 5.



Gambar 5. Daftar Pasangan Kategori Berita yang Paling Sering Tertukar (*Confused Pairs*)

Penjelasan secara kontekstual terhadap pola bias kesalahan pada Gambar 5 dapat diuraikan sebagai berikut: (1) Pola *Finance* ke *News* (9 kasus): Artikel bertopik ekonomi makro banyak membahas mengenai rilis kebijakan anggaran kementerian, birokrasi, dan regulasi pemerintah, yang mana diksi-diksi tersebut juga bertebaran sangat padat pada artikel berita umum (*News*). (2) Pola Inet ke *Finance* (9 kasus): Berita tekno*log*i di era modern kerap mengulas isu pendanaan

startup unicorn, transaksi finansial berbasis *financial technology* (fintech), investasi digital, dan performa saham perusahaan tekno*log*i global, sehingga model bias akibat tingginya bobot TF-IDF pada istilah-istilah ekonomi tersebut. (3) Pola Oto ke *Sport* (6 kasus): Artikel otomotif pada portal Detik.com sering memuat berita peluncuran motor balap baru, profil sirkuit, ulasan kompetisi balapan MotoGP, serta ajang Formula E. Istilah tersebut memiliki kedekatan semantik yang sangat erat dengan rumpun kategori olahraga (*Sport*).

4. KESIMPULAN

Penelitian pada '*NewsClassify-ID*' telah berhasil merealisasikan sistem klasifikasi kategori berita Indonesia secara otomatis dan objektif. Melalui pemanfaatan teknik *web scraping*, berhasil dihimpun *dataset* bersih sebanyak 1.199 artikel dari tujuh kategori pada portal Detik.com dengan skema pelabelan otomatis berbasis *parsing* pola *subdomain* URL. Eksperimen pembentukan representasi teks menyimpulkan bahwa kombinasi komponen 'Judul + 3 Paragraf Pertama' mampu bertindak sebagai fitur yang optimal dan efisien untuk menyajikan informasi dokumen secara ringkas. Dari hasil komparasi tiga arsitektur model, *Logistic Regression* yang dikombinasikan dengan pembobotan matriks TF-IDF mengungguli model lainnya dengan mencatatkan tingkat *Accuracy* tertinggi sebesar 86,67% dan F1-*Macro* sebesar 78,84%. Sebaliknya, arsitektur berbasis *Deep Learning Text*CNN memberikan hasil paling rendah (67,08%) akibat keterbatasan volume jumlah data latih. Analisis *confusion matrix* secara empiris membuktikan bahwa kesalahan model terkonsentrasi pada wilayah kelas yang saling beririsan konteks semantiknya, seperti Finance dengan News, serta Otomotif dengan Olahraga.

DAFTAR PUSTAKA

[1] A. A. Martha et al., "Web Scraping Dan Natural Language Processing Menggunakan Cnn Untuk Analisis Sentimen Lintas Platform Digital," vol. 10, no. 1, pp. 1621–1627,2026.

[2] M. I. Maulana, U. Hayati, T. Informatika, S. Informasi, K. Cirebon, and P. Data, "Perbandingan Algoritma Naïve Bayes Dan K-Nearest Neighbors Untuk Klasifikasi Topik Berita Pada Situs Detik . Com," vol. 8, no. 3,pp. 3733–3742, 2024.

[3] A. I. Purnamasari and I. Ali, "Analisis Sentimen Komentar Berita Detik .Com Menggunakan Algoritma Suport Vektor Machine (Svm)," vol. 8, no. 3, pp. 3175–3181, 2024.

[4] B. Bramantyo, M. Pajar, K. Putra, and N. Hendrastuty, "Klasifikasi Multikelas pada Teks Judul Berita Menggunakan Algoritma Recurrent Neural Network," vol. 3, no. 1, pp. 18–29, 2025.

[5] A. Firizkiansah, A. Muhammad, I. R. Maulana,

[6] U. S. Indonesia, and K. Bekasi, "Optimasi Klasifikasi Data Teks Menggunakan Algoritma Logistic Regression dengan TF-IDF dan SMOTE," vol. 2, no. 1, pp. 29–36, 2025.

[7] Y. A. Hafiz and E. Sudarmilah, "Implementasi Web Scraping Pada Portal Berita Online," Inisiasi, vol. 12, no. 1, pp. 55–60, Nov. 2023, doi: 10.59344/inisiasi.v12i1.120.

[8] Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, and Fitri Nurapriani, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN," J. KomtekInfo, vol. 10, pp. 1–7, 2023, doi: 10.35134/komtekinfo.v10i1.330.

[9] N. Fibriyanti Arminda, N. Sulistiyowati, and T. Nur Padilah, "Implementasi Algoritma Multinomial Naive Bayes Pada Analisis Sentimen Terhadap Ulasan Pengguna Aplikasi Brimo," JATI (Jurnal Mhs. Tek.Inform., vol. 7, no. 3, pp. 1817–1822, Nov.2023, doi: 10.36040/jati.v7i3.7012.

[10] R. Muhammad Anwar and D. V. S. Y. Sakti, "Klasifikasi Ujaran Kebencian terhadap Agensi Nijisanji menggunakan Algoritma Logistic Regression pada Tweet Berbahasa Inggris," Semin. Nas. Mhs. Fak. Teknol. Inf., vol. 3, no. 2, pp. 838–845, 2024.

[11] K. B. Berdasarkan, "Dimas, dkk, Klasifikasi Berita Berdasarkan… 201," vol. 8, no.Kommit, pp. 201–206, 2014.