INOTEK, Vol.8 Agustus 2024
ISSN: 2580-3336 (Print) / 2549-7952 (Online)
Url: https://proceeding.unpkediri.ac.id/index.php/inotek/

Klasifikasi Kategori Berita Menggunakan *Naive Bayes*

Diterima:
10 Juni 2024
Revisi:
10 Juli 2024
Terbit:
1 Agustus 2024

[1]Ikrar Nusa Bhakti, [2]Ahmad Zen Sholikhin, [3]Muhammad Abi Lukman, [4]Erna Daniati, [5]Aidina Ristyawan
[123]Fakultas Teknik dan Ilmu Komputer, Sistem Informasi, Universitas Nusantara PGRI Kediri
[1]ikrarbhakti543@gmail.com, [2]Zenahmad930@gmail.com,
[3]abilukman00@gmail.com, [4]ernadaniati@unpkediri.ac.id,
[5]aidinaristi@unpkediri.ac.id

Abstrak—Klasifikasi kategori berita merupakan salah satu aplikasi penting dalam pengolahan teks dan analisis data yang bertujuan untuk mengelompokkan artikel berita ke dalam kategori tertentu secara otomatis. Penelitian ini memanfaatkan algoritma *Naive Bayes*. Proses klasifikasi dimulai dengan pengumpulan dataset berita yang sudah dikategorikan. Dataset ini kemudian dibagi menjadi data latih dan data uji. Tahap prapemrosesan teks meliputi pembersihan data, tokenisasi, dan penghapusan stop words. Model *Naive Bayes* kemudian dilatih menggunakan data latih dan dievaluasi dengan data uji untuk mengukur kinerja model berdasarkan metrik-metrik seperti akurasi, presisi, recall. Hasil penelitian menunjukkan bahwa algoritma *Naive Bayes* mampu memberikan performa yang cukup baik dalam mengklasifikasikan berita ke dalam berbagai kategori. Dengan demikian, penelitian ini menyimpulkan bahwa *Naive Bayes* adalah metode yang layak digunakan untuk klasifikasi kategori berita, meskipun ada ruang untuk perbaikan lebih lanjut dengan teknik prapemrosesan teks yang lebih canggih dan penggunaan model pembelajaran mesin yang lebih kompleks.
Kata kunci : Kategori Berita, *Naïve Bayes*, Klasifikasi

Abstract—*News category classification is an important application in text processing and data analysis, aiming to automatically group news articles into specific categories. This study utilizes the Naive Bayes algorithm. The classification process begins with the collection of a pre-categorized news dataset. This dataset is then divided into training and testing data. The text preprocessing stage includes data cleaning, tokenization, and stop word removal. The Naive Bayes model is then trained using the training data and evaluated with the testing data to measure the model's performance based on metrics such as accuracy, precision, and recall. The results show that the Naive Bayes algorithm performs quite well in classifying news into various categories. Therefore, this study concludes that Naive Bayes is a viable method for news category classification, although there is room for further improvement with more advanced text preprocessing techniques and the use of more complex machine learning models.*
Keywords*: News Category, Naïve Bayes, Classification.*

Penulis Korespondensi:

Erna Daniati,
Sistem Informasi,
Universitas Nusantara PGRI Kediri,
Email: ernadaniati@unpkediri.ac.id
ID Orcid: [https://orcid.org/0009-0008-9471-4421]

ISSN: 2580-3336 (Print) / 2549-7952 (Online)
Url: https://proceeding.unpkediri.ac.id/index.php/inotek/

I. PENDAHULUAN

Berita di media sosial menjadi lebih populer di era internet saat ini dan menjadi cara bagi masyarakat untuk mengetahui tentang apa yang telah terjadi. Portal berita adalah salah satu sumber informasi berita yang menyampaikan berita dari berbagai sumber kepada pembaca.[1] Ketika orang lebih suka membaca berita, terutama secara online, editor dan situs portal berita harus bekerja lebih keras untuk memberi orang informasi dan berita yang baik.[2]

Portal berita biasanya menawarkan kategori berita untuk memudahkan pembaca mencari berita yang diinginkan secara cepat.[3] Namun, kategori tersebut masih diklasifikasikan secara umum, sehingga pembaca harus menyaring berita dari kategori tersebut dan mengklasifikasikannya menjadi subkategori yang lebih rinci jika mereka ingin mendapatkan kategori berita yang lebih spesifik.[4] Karena jumlah berita yang meningkat dengan cepat dan sangat mirip, membaca dan menyaringnya menjadi sulit. Oleh karena itu, proses pengklasifikasian berita sangat penting.

Klasifikasi kategori berita adalah proses otomatis yang bertujuan untuk menempatkan artikel berita ke dalam satu atau lebih kategori berdasarkan konten teksnya.[5] Proses ini tidak hanya membantu dalam pengelolaan informasi, tetapi juga meningkatkan efisiensi sistem pencarian dan rekomendasi berita. Metode pembelajaran mesin, khususnya algoritma *Naive Bayes*, telah menjadi pilihan populer untuk tugas ini karena kesederhanaan, efisiensi, dan kinerjanya yang memadai dalam berbagai aplikasi klasifikasi teks.[6]

Naive Bayes seringkali memberikan hasil yang memuaskan dalam tugas klasifikasi teks.[7]Keuntungan utama dari metode ini adalah kemampuannya untuk menangani dataset yang besar dengan cepat dan dengan sumber daya komputasi yang relatif rendah.[8] Penelitian ini akan mengeksplorasi penerapan algoritma *Naive Bayes* dalam klasifikasi kategori berita.[9] Prosesnya mencakup pengumpulan dan prapemrosesan dataset berita, pelatihan model *Naive Bayes*, serta evaluasi kinerja model menggunakan metrik-metrik seperti akurasi, presisi, dan recall.[10] Penelitian ini juga akan membahas tantangan-tantangan yang dihadapi selama proses klasifikasi dan potensi perbaikan yang dapat diterapkan untuk meningkatkan kinerja model.[11]

Dengan demikian, penelitian ini tidak hanya bertujuan untuk menunjukkan efektivitas algoritma *Naive Bayes* dalam klasifikasi kategori berita,[12] tetapi juga untuk memberikan wawasan tentang langkah-langkah praktis yang dapat diambil untuk mengoptimalkan sistem klasifikasi berita otomatis di masa depan.[13]

II. METODE PENELITIAN

Data yang digunakan dalam penelitian ini terdiri dari data sekunder yang diunduh dari Kaggle dalam format CSV. Selanjutnya, data diolah menggunakan perangkat lunak RapidMiner. Metode yang digunakan untuk klasifikasi adalah algoritma *Naive Bayes*. Terdapat 6 tahapan dalam penelitian ini

Tabel 1. Tabel tahapan penelitian

NO	Tahap	Aktivitas	Deskripsi Aktivitas
1	Rencana Penelitian	Menentukan tujuan penelitian, dan mengidentifikasi masalah	Pada tahap ini, peneliti merumuskan tujuan utama penelitian, mengidentifikasi masalah yang akan diteliti, dan merumuskan data yang akan diuji.
2	Pengumpulan Data	Mengumpulkan data statistik berita hoax yang ada di Indonesia tahun 2023 akhir	Peneliti akan mengumpulkan data yang diperlukan, yaitu data statistik kumpulan berita.
3	Pre-processing Data	Mengolah data yang diambil dari http://www.kaggle.com dan diproses ke aplikasi RapidMiner	Data yang sudah ada diproses ke aplikasi RapidMiner untuk mengetahui jumlah akurasi.
4	Implementasi *Naïve Bayes*	Menggunakan metode *naïve bayes* untuk mengklasifikasi kategori berita.	Peneliti mengimplementasikan *naïve bayes* menggunakan tool aplikasi rapidminer untuk mengklasifikasi kategori berita berdasarkan data yang sudah didapatkan.
5	Analisis Hasil	Menganalisis hasil klasifikasi *naïve bayes* dan menilai seberapa akurat hasilnya	Peneliti menganalisis hasil algoritma *naïve bayes* dalam mengklasifikasi kategori berita. Peneliti juga menilai seberapa akurat hasil klasifikasi dengan rasio 8/2.
6	Evaluation	Mengevaluasi Klasifikasi *Naïve bayes* melaui pengujian berulang	Peneliti mengevaluasi hasil analisis dengan melakukan pengujian berulang, menambah dan mengurangi jumlah data yang diuji untuk melihat seberapa besar perubahan dalam tingkat akurasi.

2.1 Klasifikasi *Naïve Bayes*

Naive Bayes adalah klasifikasi paling sederhana dan paling umum digunakan. *Naive Bayes* menghitung probabilitas kelas berdasarkan distribusi kata dalam dokumen. *Naïve Bayes* memiliki beberapa keunggulan, seperti kesederhanaan, kecepatan dan keakuratan . Adapun persamaan teorema *naive bayes* sebagai berikut:

$$(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

2.2 Confussion Matrix

Confusion Matrix merupakan metode yang dapat digunakan untuk menghitung keakuratan proses klasifikasi [14] Anda dapat menggunakan matriks konfusi untuk menganalisis seberapa baik pengklasifikasi mengenali record di kelas yang berbeda. Ini adalah tabel Confusion Matrix.

Tabel 2. Tabel Confussion Matrix

	Prediksi	
Aktual	**Positif**	**Negatif**
Positif	TP	FN
Negatif	FP	TN

2.3 Akurasi

Akurasi adalah metode pengujian yang mengukur seberapa dekat nilai yang diharapkan dengan nilai aktual. Mengidentifikasi jumlah data yang diklasifikasikan dengan benar dapat digunakan untuk memverifikasi keakuratan hasil prediksi [15].

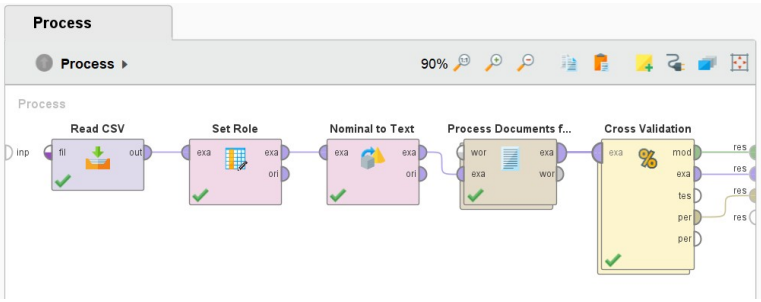
$$Accuracy = \frac{(TP+TN)}{TP+TN+FP+FN}$$

2.4 Datasets

Sumber data dari penelitian ini merupakan data sekunder, diambil dari dataset publik yang dapat diakses oleh masyarakat umum. Sumber data berasal dari kaggle.com [16] yang mana sumber tersebut menyediakan dataset sebanyak 4 kolom dan 4818 baris data tentang berita hoaxs secara factual .

III. HASIL DAN PEMBAHASAN

3.1 Text Classification Workflow

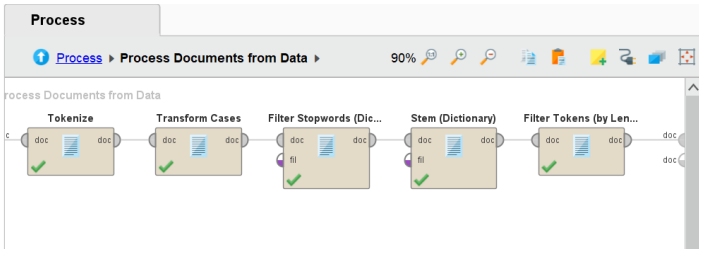


Gambar.1 *Model Text Classification Workflow*

Gambar di atas menunjukkan alur kerja pemrosesan data menggunakan RapidMiner, yang merupakan sebuah platform untuk analisis data dan machine learning. Proses ini melibatkan beberapa langkah utama untuk mempersiapkan data teks, mengonversinya ke bentuk yang dapat digunakan oleh algoritma machine learning, membangun model klasifikasi, dan mengevaluasi kinerja model tersebut. Alur kerja ini bertujuan untuk menghasilkan model klasifikasi teks yang andal dan mengevaluasi kinerjanya secara objektif menggunakan teknik validasi silang. Berikut adalah rangkuman langkah-langkah utamanya:

1. **Membaca Data (Read CSV)**: Mengimpor data dari file CSV ke dalam RapidMiner.
2. **Mengatur Peran Atribut (Set Role)**: Menentukan peran masing-masing atribut dalam data, seperti atribut mana yang akan digunakan sebagai fitur dan mana yang akan digunakan sebagai label.
3. **Mengonversi Nominal ke Teks (Nominal to Text)**: Mengubah atribut yang berbentuk kategorikal (nominal) menjadi teks untuk mempersiapkan pemrosesan teks.
4. **Memproses Dokumen dari Data (Process Documents from Data)**: Melakukan pemrosesan teks seperti tokenisasi, stemming, dan penghapusan stopwords untuk mempersiapkan teks agar dapat digunakan dalam model machine learning.
5. **Validasi Silang (Cross Validation)**: Melakukan evaluasi model dengan validasi silang untuk menilai kinerja model pada data yang belum pernah dilihat sebelumnya .

3.2 Preprocessing Data



Gambar.2 *Preprocessing Data*

Gambar di atas menunjukkan detail dari langkah Preprocessing Data yang ada dalam alur kerja RapidMiner untuk pemrosesan teks. Berikut adalah penjelasan dari setiap langkah dalam proses ini.

1. **Tokenize**

   Tokenization adalah proses memecah sekumpulan kata menjadi unit-unit yang lebih kecil yang memiliki arti tertentu, yang disebut token Dalam konteks RapidMiner, setiap token bisa diberikan nilai atau bobot tertentu berdasarkan algoritma tokenisasi yang digunakan. Misalnya, kata "aakash" bisa diberikan bobot 0.280, yang menunjukkan nilai atau kepentingan kata tersebut dalam analisis teks.

Row No.	att1	news_categ...	aachoo	aadmi	aakash	aamir	aapke	aaron
44	43	sports	0	0	0	0	0	0
45	44	sports	0	0	0	0	0	0
46	45	sports	0	0	0	0	0	0
47	46	sports	0	0	0	0	0	0
48	47	sports	0	0	0	0	0	0
49	48	sports	0	0	0.280	0	0	0
50	49	sports	0	0	0	0	0	0
51	50	world	0	0	0	0	0	0
52	51	world	0	0	0	0	0	0
53	52	world	0	0	0	0	0	0
54	53	world	0	0	0	0	0	0

Gambar.3 *Tokenization*

1. **Transform Cases**

   Langkah ini mengubah semua huruf dalam teks menjadi huruf kecil (lowercase). Transformasi ini penting untuk memastikan bahwa analisis tidak terpengaruh oleh perbedaan huruf besar dan kecil, sehingga kata-kata seperti "Data" dan "data" dianggap sama.

2. **Filter Stopwords (Dictionary)**

Langkah ini menghapus stopwords dari teks. Stopwords adalah kata-kata umum yang tidak memberikan banyak informasi penting untuk analisis, seperti "and", "the", "is", dll. Menghapus stopwords membantu mengurangi kebisingan dalam data.

3. **Stemming (Dictionary)**

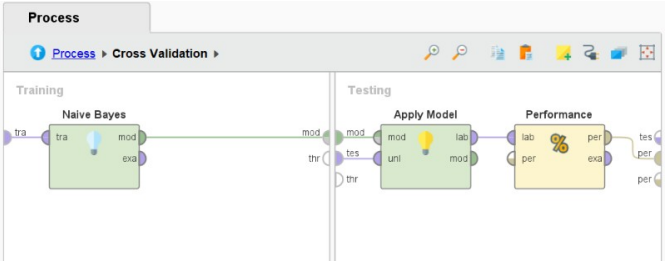
   Langkah ini melakukan stemming, yaitu mengubah kata-kata ke bentuk dasarnya (root form). Misalnya, kata "running" akan diubah menjadi "run". Stemming membantu mengurangi variasi kata-kata dan menyederhanakan analisis.

4. **Filter Tokens (by Length)**

   Langkah ini memfilter token berdasarkan panjangnya. Misalnya, token yang sangat pendek atau sangat panjang mungkin tidak relevan untuk analisis dan dapat dihapus.

Secara keseluruhan, langkah-langkah ini adalah bagian dari Preprocessing Data atau pra-pemrosesan data yang bertujuan untuk membersihkan dan menyiapkan data agar dapat digunakan dalam model machine learning dengan lebih efektif dan efisien.

3.3 Cross Validation



Gambar.4 *Di dalam Cross Validation*

Gambar di atas menunjukkan bagian dari proses **"Cross Validation"** dalam RapidMiner, yang digunakan untuk mengevaluasi model machine learning. Berikut adalah penjelasan dari setiap komponen dalam gambar tersebut:

1. ***Naive Bayes*** **(Training)**:

- Pada bagian kiri (Training), modul ini menunjukkan bahwa algoritma *Naive Bayes* digunakan untuk melatih model. Data yang digunakan untuk pelatihan (tra) dimasukkan ke dalam modul ini, dan model yang dihasilkan (mod) diteruskan ke langkah berikutnya.

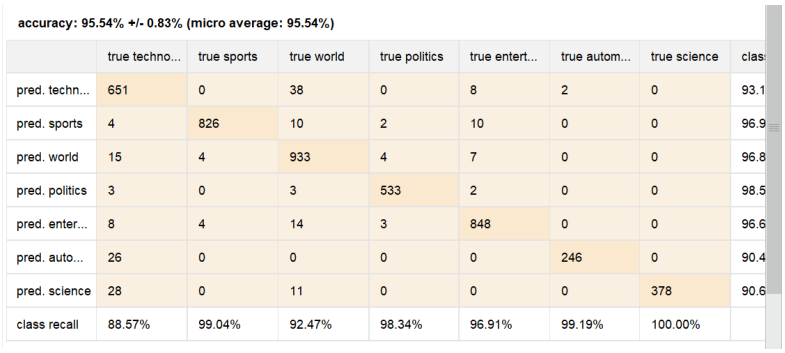
2. **Apply Model (Testing)**:

   - Pada bagian kanan (Testing), model yang telah dilatih diterapkan pada data uji (tes). Modul "Apply Model" menggunakan model yang telah dibuat (mod) dan menerapkannya pada data yang belum berlabel (unl). Output dari modul ini adalah prediksi label (lab) berdasarkan model yang diterapkan.

3. **Performance (Evaluation)**:

   - Modul ini mengevaluasi kinerja model yang telah diterapkan. Evaluasi ini dilakukan dengan membandingkan prediksi model (lab) dengan label aktual (tes) pada data uji. Hasil evaluasi memberikan metrik kinerja seperti akurasi, presisi, dan recall, yang membantu dalam menilai seberapa baik model tersebut bekerja.

Secara keseluruhan, proses ini menunjukkan bagaimana model *Naive Bayes* dilatih dan diuji menggunakan validasi silang. Validasi silang adalah teknik yang membagi data menjadi beberapa subset, melatih model pada satu subset, dan menguji model pada subset lainnya, sehingga memastikan bahwa model dievaluasi secara obyektif dan kinerjanya dapat diandalkan.



accuracy: 95.54% +/- 0.83% (micro average: 95.54%)

	true techno...	true sports	true world	true politics	true entert...	true autom...	true science	clas
pred. techn...	651	0	38	0	8	2	0	93.1
pred. sports	4	826	10	2	10	0	0	96.9
pred. world	15	4	933	4	7	0	0	96.8
pred. politics	3	0	3	533	2	0	0	98.5
pred. enter...	8	4	14	3	848	0	0	96.6
pred. auto...	26	0	0	0	0	246	0	90.4
pred. science	28	0	11	0	0	0	378	90.6
class recall	88.57%	99.04%	92.47%	98.34%	96.91%	99.19%	100.00%	

Gambar.5 *Hasil Akurasi dari Klasifikasi Berita*

Gambar ini menunjukkan hasil evaluasi model klasifikasi di RapidMiner. Model ini memiliki akurasi 95.54%, yang menunjukkan bahwa model ini dapat mengklasifikasikan data dengan cukup baik. Namun, presisi dan recall untuk beberapa kelas tidak terlalu tinggi, yang menunjukkan bahwa model ini mungkin perlu ditingkatkan.

KESIMPULAN

Kesimpulan bahwa akurasi *Naive Bayes* sebesar 95.54% menunjukkan bahwa model *naive bayes* mampu mengklasifikasikan data dengan benar untuk sekitar 95.54% kasus dalam dataset yang digunakan.Penting untuk mempertimbangkan konteks dan faktor-faktor lainnya seperti confusion matrix, precision, recall, dan mungkin metrik-metrik relevan lainnya dalam kasus seperti ini. Dalam banyak kasus, hasil yang lebih lengkap dapat diperoleh dengan melihat metrik evaluasi lain dan melakukan analisis yang lebih mendetail tentang data dan model yang digunakan. Meskipun akurasi memberikan gambaran umum tentang kinerja model, itu tidak memberikan gambaran lengkap tentang seberapa baik model memprediksi berbagai kelas atau seberapa stabil hasilnya, dalam penelitian ini dapat disimpulkan bahwa model ini dapat digunakan sebagai referensi dan dapat dilanjutkan ke penelitian dengan model prediksi yang lebih kompleks lagi.

DAFTAR PUSTAKA

[1] D. N. Chandra, G. Indrawan, and N. Sukajaya, ‘Klasifikasi Berita Lokal Radar Malang Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram’, 2016.

[2] S. Muhammad Habib, E. Haerani, S. Kurnia Gusti, S. Ramadhani, and T. H. Informatika UIN Sultan Syarif Kasim Riau Jl Soebrantas, ‘Klasifikasi Berita Menggunakan Metode Naïve Bayes Classifier’, *Jurnal Nasional Komputasi dan Teknologi Informasi*, vol. 5, no. 2, 2022.

[3] R. Firmansyah, ‘Web Klarifikasi Berita Untuk Meminimalisir Penyebaran Berita Hoax’, *JURNAL INFORMATIKA*, vol. 4, no. 2, 2017.

[4] F. Prasetya and F. Ferdiansyah, ‘Analisis Data Mining Klasifikasi Berita Hoax COVID 19 Menggunakan Algoritma Naive Bayes’, *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 4, no. 1, p. 132, Sep. 2022, doi: 10.30865/json.v4i1.4852.

[5] D. Santi, J. Nangi, and N. Ransi, ‘Implementasi Naïve bayes Clasifier dalam Klasifikasi Jenis Berita’, *Foristek*, vol. 10, no. 1, Mar. 2020, doi: 10.54757/fs.v10i1.52.

[6] M. N. Randhika, J. C. Young, A. Suryadibrata, and H. Mandala, ‘Implementasi Algoritma Complement dan Multinomial Naïve Bayes Classifier Pada Klasifikasi Kategori Berita Media Online’, *Ultimatics : Jurnal Teknik Informatika*, vol. 13, no. 1, 2021.

[7] Y. D. Pramudita, S. S. Putro, and N. Makhmud, 'Klasifikasi Berita Olahraga Menggunakan Metode Naïve Bayes dengan Enhanced Confix Stripping Stemmer', *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 3, pp. 269–276, Aug. 2018, doi: 10.25126/jtiik.201853810.

[8] S. Parsaoran Tamba, A. Laia, Y. Kristian Butar Butar, and F. Sains dan Teknologi, 'PENERAPAN DATA MINING UNTUK KLASIFIKASI BERITA HOAX MENGGUNAKAN ALGORITMA NAIVE BAYES', *Jurnal TEKINKOM*, vol. 6, no. 2, p. 2023, doi: 10.37600/tekinkom.v6i2.922.

[9] S. Sukriadi, I. Ismail, and A. M. Andzar, 'Penerapan Text Mining Dalam Klasifikasi Judul Skripsi Yang Diusulkan Mahasiswa Menggunakan Metode Naïve Bayes', *Jurnal Ilmiah Sistem Informasi dan Teknik Informatika (JISTI)*, vol. 6, no. 2, pp. 184–196, Oct. 2023, doi: 10.57093/jisti.v6i2.174.

[10] E. Triawan, N. Suarna, and A. Rinaldi Dikananda, 'KLASIFIKASI TIPE PENYERANG SEPAK BOLA LIGA INGGRIS BERDASARKAN DATA STATISTIK PEMAIN MENGGUNAKAN METODE NAIVE BAYES', 2024.

[11] S. K. Dirjen *et al.*, 'Terakreditasi SINTA Peringkat 2 Klasifikasi Berita Menggunakan Algoritma Naive Bayes Classifer Dengan Seleksi Fitur Dan Boosting', *masa berlaku mulai*, vol. 1, no. 3, pp. 227–232, 2017.

[12] S. W. Ritonga, . Y., M. Fikry, and E. P. Cynthia, 'Klasifikasi Sentimen Masyarakat di Twitter terhadap Ganjar Pranowo dengan Metode Naïve Bayes Classifier', *Building of Informatics, Technology and Science (BITS)*, vol. 5, no. 1, Jun. 2023, doi: 10.47065/bits.v5i1.3535.

[13] R. Rakhmat Sani, Y. Ayu Pratiwi, S. Winarno, E. Devi Udayanti, and dan Farrikh Al Zami, 'Analisis Perbandingan Algoritma Naive Bayes Classifier dan Support Vector Machine untuk Klasifikasi Hoax pada Berita Online Indonesia', 2022.

[14] S. Parsaoran Tamba, A. Laia, Y. Kristian Butar Butar, and F. Sains dan Teknologi, 'PENERAPAN DATA MINING UNTUK KLASIFIKASI BERITA HOAX MENGGUNAKAN ALGORITMA NAIVE BAYES', *Jurnal TEKINKOM*, vol. 6, no. 2, p. 2023, doi: 10.37600/tekinkom.v6i2.922.

[15] S. K. Dirjen *et al.*, 'Terakreditasi SINTA Peringkat 2 Klasifikasi Berita Menggunakan Algoritma Naive Bayes Classifer Dengan Seleksi Fitur Dan Boosting', *masa berlaku mulai*, vol. 1, no. 3, pp. 227–232, 2017.

[16] "Kaggle, News Classification "

https://www.kaggle.com/datasets/kishanyadav/inshort-news

accessed, June 03, 2024