e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Vol. 9 No. 1, April, 2025, **Hal. 60-68**
DOI: 10.29408/edumatic.v9i1.28987

Performa Logistic Regression dan Naive Bayes dalam Klasifikasi Berita Hoax di Indonesia

Okta Nur Cahyani [1,*], Fikri Budiman [1]

[1] Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Indonesia
[*] Correspondence: 111202113600@mhs.dinus.ac.id

Received: 23 Desember 2024 | Revised: 29 Desember 2024 | Accepted: 24 Januari 2025 | Published: 10 April 2025

Abstrak
Penyebaran informasi yang tidak benar menjadi tantangan utama dalam masyarakat Indonesia, dengan 2.484 kasus yang terdaftar pada tahun 2022. Hal ini menunjukkan pentingnya pengembangan sistem yang mampu secara efektif mengidentifikasi dan menyaring berita *hoax*. Penelitian ini bertujuan untuk mengembangkan model deteksi berita *hoax* yang lebih tepat dengan memanfaatkan *logistic regression* yang dioptimalkan melalui *grid search* dan *oversampling* untuk menangani ketidakseimbangan data. Fokus utama penelitian ini adalah untuk meningkatkan performa model dalam mendeteksi berita *hoax* pada dataset yang tidak seimbang. Dataset yang digunakan adalah *false news* Indonesia yang terdiri dari 4. 231 data, dengan dua kategori: valid (3465 data) dan *hoax* (766 data). Proses dilakukan melalui langkah-langkah seperti stemming, penghapusan stopword, dan normalisasi teks menggunakan TF-IDF. *Random oversampling* diterapkan untuk menyeimbangkan data antara kelas *hoax* dan valid, serta optimalisasi parameter menggunakan *grid search* untuk meningkatkan kinerja model. Hasil penelitian menunjukkan bahwa *logistic regression* yang dioptimalkan memberikan *accuracy* tertinggi sebesar 93%, mengungguli *naive bayes* dengan *accuracy* 86%. Hasil ini menunjukkan bahwa model deteksi berita *hoax* yang dikembangkan dapat dimanfaatkan untuk memperbaiki sistem pemantauan berita di media sosial, serta meningkatkan literasi digital masyarakat Indonesia.

Kata kunci: berita *hoax; grid search; logistic regression*; *naive bayes*

Abstract
The spread of false information has become a major challenge in Indonesian society, with 2,484 cases recorded in 2022. This highlights the importance of developing a system that can effectively identify and filter out fake news. This research aims to develop a more accurate fake news detection model by applying logistic regression, which is optimized by grid search and oversampling to overcome data imbalance. The main focus of this research is to improve the performance of the model in detecting fake news on unbalanced datasets. The dataset used is the Indonesian Fake News dataset, which consists of 4,231 entries with two categories: valid (3,465 entries) and hoax (766 entries). Preprocessing steps include stemming, stopword removal, and text normalization using TF-IDF. Random oversampling was applied to balance the data between hoax and valid classes, and parameter optimization was performed using grid search to improve model performance. The results show that the optimized logistic regression achieved the highest accuracy of 93%, surpassing naive bayes, which achieved 86% accuracy. These findings suggest that the developed fake news detection model can be used to improve the social media news monitoring system, and increase digital literacy among Indonesians.

Keywords: *fake news; grid search; logistic regression; naive bayes*

PENDAHULUAN

Berita *hoax* adalah informasi yang salah dan menyesatkan yang dapat *menimbulkan* kebencian bagi pembaca (Putri et al., 2020; Roy & Junaidi, 2020). Di era digital yang semakin

berkembang, kemajuan dalam teknologi informasi telah membuat distribusi dan akses informasi menjadi lebih mudah melalui berbagai *platform* media sosial. Namun, kemajuan ini juga membuka celah bagi penyebaran berita *hoax* yang dapat membahayakan masyarakat, terutama dalam hal kesehatan, politik, dan bencana (Pardede & Ibrahim, 2020). Menurut Cambridge Dictionary, *hoax* merujuk pada tindakan penipuan atau tipuan yang dirancang untuk membingungkan orang lain (Kurniawan & Mustikasari, 2022).

Penyebaran berita *hoax* di Indonesia semakin mengkhawatirkan, terutama di media sosial yang memungkinkan informasi menyebar dengan cepat dari satu akun ke akun lainnya. Laporan dari Kementerian Komunikasi dan Informatika (KOMINFO) menunjukkan bahwa 2.484 kasus berita *hoax* tersebar melalui media sosial pada tahun 2022, terutama yang berkaitan dengan politik, kesehatan, dan bencana. Selama pandemi COVID-19, berita h*oax* tentang pengobatan dan konspirasi vaksin menghambat upaya pemerintah untuk meningkatkan jumlah vaksinasi (Purnajaya & Pernando, 2023; Ropikoh et al., 2021). Kondisi ini menunjukan urgensi untuk mengembangkan sistem deteksi yang efektif dalam mengidentifikasi berita *hoax*.

Penelitian ini mengusulkan solusi untuk mengatasi tantangan dalam mendeteksi berita *hoax*. Dua pendekatan utama yang digunakan adalah random *oversampling* dan *grid search*. Random *oversampling* memecahkan masalah ketidakseimbangan data dengan meningkatkan jumlah sampel kelas minoritas, yaitu Berita *hoax*. Hal ini memungkinkan model mempelajari karakteristik berita *hoax* dengan lebih baik dan akhirnya mengidentifikasinya dengan lebih akurat. Sementara itu penelitian ini juga menerapkan *grid search* untuk mengoptimalkan *hyperparameter* model, yang memungkinkan membuat kombinasi parameter terbaik. sehingga dapat menghasilkan kombinasi parameter terbaik yang dapat meningkatkan kinerja model pada metrik evaluasi yang lebih komprehensif, seperti *recall, precision, accuracy,* dan *f1-score* (Fauzi & Yunial, 2022; Matin, 2023).

Pendekatan ini didasarkan pada teori bahwa *random oversampling* dapat mengoreksi ketidakseimbangan kelas yang sering terjadi dalam kumpulan data berita *hoax*. Seringkali, kumpulan data yang digunakan untuk melatih model berita *hoax* didominasi oleh kelas mayoritas, yaitu valid. Oleh karena itu, model tersebut cenderung lebih baik dalam mengklasifikasikan pesan yang valid. Dengan meningkatkan data pada kelas *hoax*, model dapat mempelajari lebih banyak fitur *hoax* menggunakan *random oversampling*, yang seharusnya membuat model lebih efisien dalam mendeteksi *hoax*. Selain itu, dapat menggunakan fitur *grid search* untuk menemukan konfigurasi *hyperparameter* yang optimal. Hal ini berdampak langsung pada kinerja model, menghasilkan klasifikasi yang lebih akurat dan tepat (Ananda et al., 2024; Fitriani et al., 2021; Nurrokhman, 2023).

Penelitian sebelumnya yang dilakukan oleh Ruise et al. (2023) yang membandingkan tiga algoritma *machine learning* SVM, *logistic regression*, dan *random forest* untuk mendeteksi berita *hoax* di Twitter, terungkap bahwa walaupun tingkat *accuracy* model cukup tinggi, metrik evaluasi lain seperti *presisi* dan *recall* tidak diambil dalam pertimbangan. Hal ini menjadi isu karena akurasi yang tinggi dapat dipicu oleh dominasi kelas mayoritas (berita *valid*), sehingga tidak mencerminkan kemampuan model dalam mengidentifikasi berita *hoax*. Selain itu, penelitian tersebut tidak menerapkan teknik pengelolaan ketidakseimbangan data, seperti *oversampling*, yang bisa mengarah pada bias terhadap kelas yang lebih banyak.

Beberapa tren terbaru menunjukkan betapa pentingnya teknologi informasi untuk mendeteksi berita *hoax*, menunjukkan bahwa algoritma *logistic regression* dan *naive bayes* berfungsi dengan baik untuk membedakan berita *hoax* dari berita asli (Ahmed Arafa et al., 2022; Ardiansyah et al., 2021). Untuk mengatasi masalah penelitin sebelumnya, penelitian ini menggunakan metrik evaluasi seperti *accuracy, precision, recall*, dan *f1-score* dengan fokus pada *recall* untuk memastikan bahwa model dapat mendeteksi berita *hoax* secara optimal (Ramadhan et al., 2022; Sani et al., 2022; Zhafira et al., 2021). Metode *random oversampling* diterapkan untuk menyeimbangkan data, meningkatkan jumlah data pada kelas *hoax*, sehingga

model dapat mempelajari pola berita *hoax* dengan lebih efektif dan meningkatkan kinerja *recall* serta *f1-score*. Selain itu, *grid search* diterapkan untuk mengoptimalkan *hyperparameter* pada *logistic regression* dan *naive bayes* guna meningkatkan performa model. *Logistic regression* dipilih karena sesuai untuk data berdimensi tinggi, sementara *naive bayes* efisien dalam menangani fitur dengan distribusi yang sederhana (Armansyah & Ramli, 2022; Hendrawan et al., 2022; Lindawati et al., 2023). Langkah ini juga menyempurnakan hasil temuan sebelumnya, dengan fokus pada peningkatan kinerja model dalam klasifikasi berita *hoax* di Indonesia (Afrizal et al., 2020; Gifari et al., 2022; Tanggraeni & Sitokdana, 2022).

Penelitian Muhabatin et al. (2021) berupaya meningkatkan *accuracy* klasifikasi berita *hoax* menggunakan *naïve bayes* yang dioptimalkan dengan *Particle Swarm Optimization* (PSO). Namun, ada beberapa kelemahan seperti dataset yang digunakan terbatas, hanya 110 data, metode validasi seperti *cross-validation k-fold* tidak dijelaskan, dan algoritma pembanding lain, seperti SVM atau *logistic regression* tidak disertakan. Meskipun ada peningkatan *accuracy* sebesar 18,18%, metrik seperti ketepatan, *recall*, dan *f1-score* tidak diperiksa. terutama penting untuk dataset tidak seimbang. Untuk hasil yang lebih andal, disarankan penelitian lebih lanjut dengan dataset yang lebih besar, validasi yang lebih kuat, dan perbandingan metode klasifikasi yang berbeda.

Penelitian ini bertujuan untuk mengembangkan model deteksi berita *hoax* yang lebih akurat menggunakan *logistic regression* yang dioptimalkan melalui *grid search* dan *oversampling*. Fokus penelitian ini adalah pada peningkatan kemampuan model dalam mendeteksi berita *hoax*, terutama pada kumpulan data yang tidak seimbang. Hasil penelitian ini dapat digunakan untuk mengembangkan sistem pemantauan media sosial yang dapat secara otomatis mendeteksi dan mengurangi penyebaran berita *hoax*. Selain itu, model ini juga dapat berfungsi sebagai landasan untuk meningkatkan literasi digital masyarakat, termasuk mengedukasi mereka tentang pola umum dalam berita *hoax* dan peningnya berpikir kritis saat dalam menerima informasi. Dengan pendekatan ini, penelitian ini tidak hanya menambah wawasan akademik, tetapi juga memberikan dampak langsung bagi masyarakat menerima informasi. Pendekatan ini memastikan bahwa penelitian tidak hanya memberikan wawasan akademis tetapi juga.

METODE

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan eksperimen komputasi dan pembelajaran mesin (*machine learning*). Metode yang diterapkan mencakup algoritma *logistic* regression dan *naive bayes* untuk mendeteksi berita *hoax* dalam dataset *false news* (*hoax*) Indonesia dari Kaggle yang terdiri dari 4231 data, dengan dua kategori valid (3465) dan *hoax* (766). *Preprocessing* meliputi pembersihan data, penanganan data yang hilang, penghapusan kolom yang tidak relevan, serta pelaksanaan label encoding, stemming, tokenisasi, dan penghapusan *stopword*. TF-IDF diterapkan untuk representasi numerik teks, yang sesuai untuk dataset dengan skala sedang.

Random oversampling digunakan untuk menyeimbangkan kelas, sementara *grid search* mengoptimalkan *hyperparameter* guna mengurangi risiko *overfitting*. Data dipecah dengan rasio 80:20 untuk pelatihan dan pengujian. Walaupun TF-IDF efisien untuk dataset skala sedang, metode lain seperti Word2Vec dan BERT lebih sesuai untuk dataset besar karena mampu menangkap konteks semantik kata dengan lebih mendalam.

logistic regression dan *naive bayes* dipilih karena efektif untuk pengklasifikasian teks pada dataset dengan skala menengah dan dapat menangani ketidakseimbangan kelas. *Logistic regression* memanfaatkan fungsi logistik untuk mengaitkan input dengan probabilitas kelas yang ditargetkan, sedangkan *naive bayes* menggunakan teorema bayes untuk mendapatkan independensi fitur. *Grid search* diterapkan untuk mengoptimalkan parameter model. Parameter untuk masing-masing model, yang dapat dilihat dari tabel 1 dan 2

Tabel 1. Parameter *logistic regression*

Parameter	Value
C	[0.1, 1, 10, 100]
Penalty	[l1, l2]
Solver	[liblinear, lbfgs, newton-cg]
Max_iter	[100, 200, 300]

Tabel 1 adalah parameter yang digunakan untuk mengoptimalkan model *machine learning*, khususnya pada *logistic regression*. Parameter C menunjukkan tingkat regulasi dengan nilai [0. 1, 1, 10, 100], di mana nilai yang lebih kecil memberikan regulasi yang lebih kuat. Penalty mengacu pada jenis regulasi, yaitu l1 (Lasso) dan l2 (Ridge). *Solver* merupakan algoritma optimasi yang digunakan, termasuk liblinear, lbfgs, dan newton-cg. Dan Max_iter menetapkan jumlah iterasi maksimum untuk konvergensi, dengan pilihan [100, 200, 300]. Parameter-parameter ini biasanya diuji dalam *grid search* untuk menemukan kombinasi terbaik.

Tabel 2. Parameter *naive bayes*

Parameter	**Value**
Alpha	[0.1, 0.5, 1.0, 2.0]
Fit prior	[True, False]

Selanjutnya pada tabel 2 menjelaskan parameter *alpha* mengatur tingkat *smoothing* dengan nilai [0. 1, 0. 5, 1. 0, 2. 0], di mana nilai yang lebih tinggi memberikan *smoothing* yang lebih kuat untuk mencegah pembagian oleh nol. *Fit prior* adalah pilihan untuk menentukan apakah probabilitas prior kelas dihitung dari data (*true*) atau diabaikan (*false*). Parameter ini diuji untuk memperoleh performa model yang terbaik.

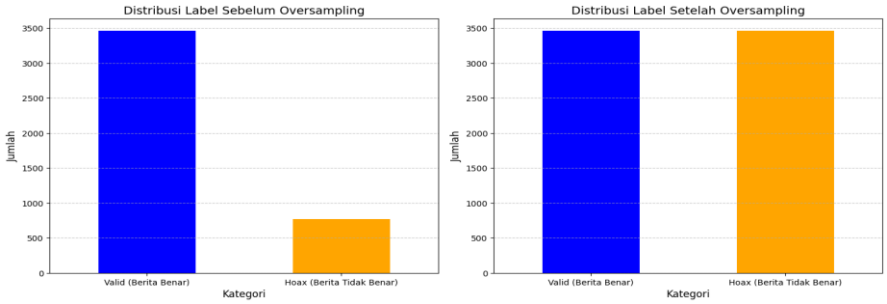
Upaya untuk meningkatkan performa model dilakukan dengan menerapkan algoritma *logistic regression* dan *naive bayes* yang dievaluasi menggunakan beberapa metrik penting. Akurasi digunakan untuk mengukur persentase prediksi yang benar dari total data uji, memberikan gambaran umum tentang seberapa baik model berfungsi secara keseluruhan. *precision* menilai akurasi prediksi positif, yang sangat penting untuk mengurangi kesalahan prediksi terhadap berita *hoax*. *Recall* mengukur kemampuan model dalam menangkap seluruh data positif yang sebenar-benarnya, sehingga relevan untuk memastikan semua berita *hoax* terdeteksi dengan baik. *F1-score* menggabungkan *precision* dan *recall* untuk mengevaluasi keseimbangan antara kedua metrik tersebut, terutama pada dataset yang tidak seimbang. Dengan menggunakan metrik-metrik ini, model terbaik dapat diidentifikasi berdasarkan kemampuannya dalam memberikan hasil yang akurat dan dapat diandalkan.

HASIL DAN PEMBAHASAN

Hasil

Dataset yang digunakan dalam penelitian ini diperoleh dari situs Kaggle yang berjudul Indonesia *False News* (*Hoax*). Dataset ini mengandung total 4.231 data yang terbagi menjadi dua kategori, yaitu *valid* dan *hoax*. Sebelum dilakukan *oversampling*, distribusi label pada dataset menunjukkan ketidakseimbangan, di mana terdapat 3.465 data untuk label *valid* dan hanya 766 data untuk label *hoax*. Ketidakseimbangan ini dapat memengaruhi kinerja model. Sehingga dilakukan *oversampling* untuk menyamakan jumlah data di kedua kategori agar model belajar lebih baik. *Proses preprocessing* meliputi penggabungan kolom teks untuk mempermudah proses pengolahan, penghapusan kolom yang tidak relevan serta penormalan teks dengan menghilangkan tanda baca, mengubah semua huruf menjadi lebih kecil, dan

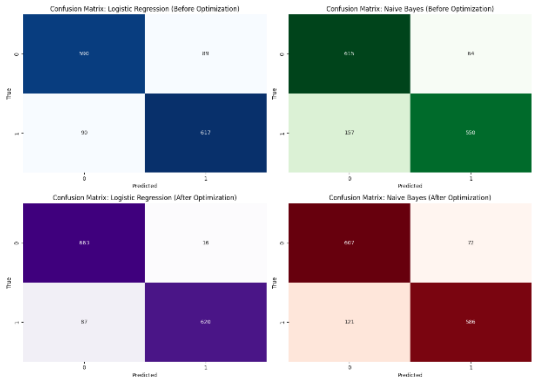
menghilangkan karakter khusus. Data teks diubah menjadi data numerik menggunakan metode *TF-IDF* yang memberikan bobot pada kata berdasarkan frekuensi kemunculannya. Teknik ini membantu meningkatkan performa model dalam mendeteksi berita.



Gambar 1. Hasil *oversampling*

Pada gambar 1 dapat dilihat hasil dari upaya mengatasi ketidakseimbangan data dalam dataset ini, dilakukan *random oversampling* yang membantu model berlajar dengan lebih baik dan meningkatkan akurasi prediksi, khususnya pada kategori *hoax* yang sebelumnya kurang terwakili. Sebelum *oversampling* data kategori *valid* lebih banyak dibandingkan kategori *hoax*, namun setelah keduanya seimbang akurasi model juga meningkat.

Penelitian ini membandingkan kinerja *logistic regression* dan *multinomial naive bayes* dalam klasifikasi berita *hoax* dan *valid*. *Logistic regression* menggunakan fungsi logistic untuk memprediksi kemungkinan suatu kategori, sedangkan *multinomial naive bayes* menghitung kemungkinan kategori berdasarkan distribusi kata dalam data. Dengan menggunakan *grid search*, keduanya dioptimalkan untuk menemukan parameter terbaik dan meningkatkan *accuracy*, *presisi*, *recall*, dan *f1-score*. Hasil optimasi menunjukkan peningkatan *logistic regression*, dengan *False Positives* (FP) turun dari 89 menjadi 16, *False Negatives* (FN) turun dari 90 menjadi 87, dan *True Negatives* (TN) meningkat dari 590 menjadi 663. Sementara itu, Multinomial *Naive Bayes* menunjukkan peningkatan yang lebih sedikit, dengan FP meningkat dari 64 menjadi 72. Secara keseluruhan, *logistic regression* ternyata lebih baik untuk mengklasifikasikan berita *hoax* dan *valid* karena memiliki tingkat kesalahan yang lebih rendah dan kinerja yang lebih konsisten. Hasil dari percobaan ini dapat dilihat dari gambar 2 dan tabel 3.



Gambar 2. *Confusion matrix*

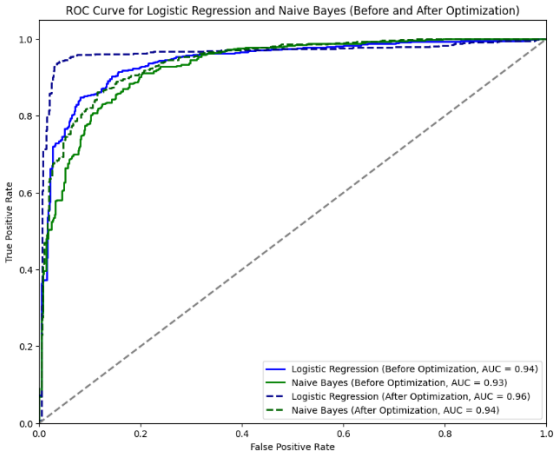
Tabel 3 menunjukkan bahwa model *logistic regression* tanpa menerapkan pencarian grid mencapai performa terbaik pada metrik *accuracy*, *precision*, *recall* dan *f1-score* dengan nilai 87,00%. Sebaliknya, model *naïve bayes* memiliki performa yang sedikit lebih rendah, yaitu 84,00% pada semua metrik ini. Setelah optimasi menggunakan pencarian grid, model *logistic regression* secara signifikan meningkatkan nilai semua metrik yang diukur sebesar 93,00%.

Sebaliknya, *naive bayes* hanya mengalami sedikit peningkatan dengan nilai akhir 86,00%. Secara keseluruhan, model *logistic regression* yang dioptimalkan pencarian grid menunjukkan hasil yang lebih baik daripada model *naive bayes*. Hal ini disebabkan oleh *accuracy* dan *recall* yang tinggi dari model ini, yang memungkinkan untuk meminimalkan kesalahan dan mendeteksi *hoax* secara luas. Hal Ini mendukung tujuan penelitian untuk mengembangkan model klasifikasi berita *hoax* yang dapat diandalkan untuk mengurangi penyebaran berita *hoax* di platform.

Tabel 3. Hasil klasifikasi algoritma

Model	***Logistic Regression***	***Naive Bayes***	***Logistic Regression + Grid Search***	***Naive Bayes + Grid Search***
Accuracy	87,00%	84,00%	93,00%	86,00%
Precision	87,00%	85,00%	93,00%	86,00%
Recall	87,00%	84,00%	93,00%	86,00%
F1-Score	87,00%	84,00%	93,00%	86,00%

Gambar 3 menggambarkan kurva *Receiver Operating Characteristic* (*ROC)* yang digunakan untuk menilai kinerja model klasifikasi *logistic regression* dan *naive bayes*, baik sebelum maupun setelah dilakukan optimasi. Kurva *ROC* menunjukkan hubungan antara False *Positive Rate* (FPR) yang berada di sumbu horizontal dan *True Positive Rate* (TPR) yang berada di sumbu vertikal. Dari gambar, terlihat bahwa model *logistic regression* memiliki nilai *Area Under the Curve* (AUC) sebesar 0.94 sebelum proses optimasi dan meningkat menjadi 0,96 setelah optimasi, menandakan adanya peningkatan kinerja yang signifikan. Di sisi lain, model *naive bayes* memiliki nilai *AUC* sebesar 0,93 sebelum optimasi dan meningkat sedikit menjadi 0,94 setelah optimasi. Garis putus-putus menggambarkan hasil setelah optimasi, dan terlihat bahwa *logistic regression* setelah optimasi menunjukkan performa terbaik dengan kurva yang lebih mendekati sudut kiri atas grafik. Hal ini menunjukkan bahwa optimasi memberikan dampak yang lebih signifikan pada *logistic regression* jika dibandingkan dengan *naive bayes*, dengan *logistic regression* setelah optimasi menjadi model dengan kinerja terbaik berdasarkan nilai *AUC*.



Gambar 3. Kurva *roc*

Pembahasan

Penelitian ini menggunakan dataset Indonesia False News (*Hoax*) yang memiliki ketidakseimbangan kelas, di mana terdapat lebih banyak data valid dibandingkan *hoax*. Untuk

mengatasi masalah ini, teknik *random oversampling* diterapkan untuk meningkatkan representasi data *hoax*. Dua algoritma, *logistic regression* dan *naive bayes*, digunakan dan dioptimalkan dengan *grid search* untuk menemukan kombinasi *hyperparameter* terbaik. Hasil penelitian menunjukkan bahwa *logistic regression* memberikan akurasi yang lebih tinggi daripada *naive bayes*. Sebelum dilakukan optimasi, *accuracy logistic regression* adalah 87%, sementara *naive bayes* hanya mencapai 84%. Setelah dilakukan optimasi, *logistic regression* meningkat menjadi 93%, sedangkan *naive bayes* hanya mengalami peningkatan kecil menjadi 86%. Hal ini menunjukkan bahwa *logistic regression* lebih efektif dalam menangani data yang lebih kompleks. Evaluasi yang menggunakan *AUC* dan *ROC* menunjukkan bahwa *logistic regression* tampil lebih baik, dengan *AUC* yang meningkat dari 0.94 menjadi 0.96 setelah dilakukan optimasi. Di sisi lain, *naive bayes* hanya mengalami sedikit peningkatan dari 0.93 menjadi 0.94, yang menunjukkan adanya keterbatasan dalam mengatasi data dengan hubungan fitur yang kompleks. Hasil penelitian ini mengindikasikan bahwa *logistic regression* lebih baik dalam klasifikasi berita *hoax* karena kemampuannya untuk mengelola data teks dengan fitur yang saling terkait, sementara *naive bayes* memiliki batasan dalam hal ini.

Penelitian ini memberikan pendekatan yang lebih lengkap dibandingkan penelitian sebelumnya yang hanya mengukur akurasi (*accuracy*) tanpa mempertimbangkan *recall* dan *f1-score*, yang penting untuk dataset tidak seimbang (Ruise et al., 2023). Dengan mengadopsi *random oversampling* dan optimasi parameter, model dalam penelitian ini menunjukkan peningkatan signifikan dalam kinerja, menjadikannya lebih andal dalam menangani data yang tidak seimbang. Hasil ini menegaskan bahwa pendekatan berbasis memori cocok untuk mendeteksi berita *hoax*. Di sisi lain, penelitian Muhabatin et al. (2021) berupaya meningkatkan akurasi klasifikasi berita *hoax* menggunakan *naïve bayes* yang dioptimalkan dengan PSO.

Meskipun penelitian mereka mencatat peningkatan akurasi sebesar 18,18%, terdapat beberapa kelemahan seperti dataset yang sangat terbatas (hanya 110 data), tidak adanya validasi seperti *k-fold cross-validation*, serta tidak dilibatkannya algoritma pembanding lain seperti SVM atau *logistic regression*. Selain itu, metrik penting seperti *precision*, *recall*, dan *f1-score* juga tidak dianalisis, sehingga hasilnya kurang mendalam dan sulit diterapkan pada dataset yang lebih besar. Dengan metode ini, model menjadi lebih andal dalam menangani data yang tidak seimbang. Hasil penelitian ini menunjukkan bahwa *random oversampling* dan optimasi parameter secara signifikan meningkatkan kinerja model. Hal ini juga menegaskan bahwa pendekatan berbasis memori cocok untuk mendeteksi berita *hoax*. Penelitian selanjutnya dapat dikembangkan dengan mengeksplorasi algoritma yang lebih kompleks seperti BERT atau menggunakan data yang lebih besar untuk meningkatkan *accuracy* dan kemampuan generalisasi model.

SIMPULAN

Penelitian ini fokus dengan pentingnya penerapan teknik dan metode pengoptimalan parameter untuk mengatasi ketidakseimbangan data dan meningkatkan kinerja model dalam klasifikasi berita *hoax*. Pendekatan ini menunjukkan efektivitas dalam mendeteksi kategori minoritas dan berpotensi digunakan dalam sistem penyaringan pesan otomatis pada *platform* digital. Sistem yang lebih akurat dapat membantu meminimalkan dampak negatif penyebaran informasi yang salah. Namun, keterbatasan dataset yang hanya mencakup berita *hoax* berbahasa Indonesia mengindikasikan perlunya studi lanjutan dengan dataset multibahasa dan struktur yang lebih kompleks guna memastikan generalisasi hasil yang lebih luas.

.

REFERENSI

Afrizal, S., Irmanda, H. N., Falih, N., & Isnainiyah, I. N. (2020). Implementasi Metode Naïve Bayes untuk Analisis Sentimen Warga Jakarta Terhadap. *Informatik: Jurnal Ilmu Komputer*, *15*(3), 157–168. https://doi.org/10.52958/iftk.v15i3.1454

Ahmed Arafa, A. H., Radad, M., Badawy, M. M., & El-Fishawy, N. (2022). Logistic regression hyperparameter optimization for cancer classification. *Menoufia Journal of Electronic Engineering Research*, *31*(1), 1-8. https://doi.org/10.21608/mjeer.2021.70512.1034

Ananda, I. K., Fanani, A. Z., Setiawan, D., & Wicaksono, D. F. (2024). Penerapan Random Oversampling dan Algoritma Boosting untuk Memprediksi Kualitas Buah Jeruk. *Edumatic: Jurnal Pendidikan Informatika*, *8*(1), 282–289. https://doi.org/10.29408/edumatic.v8i1.25836

Ardiansyah, M., Sunyoto, A., & Luthfi, E. T. (2021). Analisis Perbandingan Akurasi Algoritma Naïve Bayes dan C4.5 untuk Klasifikasi Diabetes. *Edumatic: Jurnal Pendidikan Informatika*, *5*(2), 147–156. https://doi.org/10.29408/edumatic.v5i2.3424

Armansyah, A., & Ramli, R. K. (2022). Model prediksi kelulusan mahasiswa tepat waktu dengan metode Naïve Bayes. *Edumatic: Jurnal Pendidikan Informatika*, *6*(1), 1-10. https://doi.org/10.29408/edumatic.v6i1.4789

Fauzi, A., & Yunial, A. H. (2022). Optimasi Algoritma Klasifikasi Naive Bayes, Decision Tree, K – Nearest Neighbor, dan Random Forest menggunakan Algoritma Particle Swarm Optimization pada Diabetes Dataset. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *8*(3), 470–481. https://doi.org/10.26418/jp.v8i3.56656

Fitriani, R. D., Yasin, H., & Tarno, T. (2021). Penanganan Klasifikasi Kelas Data Tidak Seimbang Dengan Random Oversampling Pada Naive Bayes(Studi Kasus: Status Peserta KB IUD di Kabupaten Kendal). *Jurnal Gaussian*, *10*(1), 11–20. https://doi.org/10.14710/j.gauss.v10i1.30243

Gifari, O. I., Adha, Muh., Freddy, F., & Durrand, F. F. S. (2022). Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine. *Journal of Information Technology*, *2*(1), 36–40. https://doi.org/10.46229/jifotech.v2i1.330

Hendrawan, I. R., Utami, E., & Hartanto, A. D. (2022). Comparison of Naïve Bayes Algorithm and XGBoost on Local Product Review Text Classification. *Edumatic: Jurnal Pendidikan Informatika*, *6*(1), 143-149. https://doi.org/10.29408/edumatic.v6i1.5613

Kurniawan, A. A., & Mustikasari, M. (2022). Evaluasi Kinerja MLLIB APACHE SPARK pada Klasifikasi Berita Palsu dalam Bahasa Indonesia. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *9*(3), 489-500. https://doi.org/10.25126/jtiik.2022923538

Lindawati, L., Fadhli, M., & Wardana, A. S. (2023). Optimasi Gaussian Naïve Bayes dengan Hyperparameter Tuning dan Univariate Feature Selection dalam Prediksi Cuaca. *Edumatic: Jurnal Pendidikan Informatika*, *7*(2), 237-246. https://doi.org/10.29408/edumatic.v7i2.21179

Matin, I. M. M. (2023). Hyperparameter Tuning Menggunakan GridsearchCV pada Random Forest untuk Deteksi Malware. *MULTINETICS*, *9*(1), 43–50. https://doi.org/10.32722/multinetics.v9i1.5578

Muhabatin, H., Prabowo, C., Ali, I., Lukman Rohmat, C., Rizki Amalia, D., sitasi, C., & Rizki, D. (2021). Klasifikasi Berita Hoax Menggunakan Algoritma Naïve Bayes Berbasis PSO. *Informatics for Educators and Professionals*, *5*(2), 156–165. https://doi.org/10.51211/itbi.v5i2.1531

Nurrokhman, M. Z. (2023). Perbandingan Algoritma Support Vector Machine dan Neural Network untuk Klasifikasi Penyakit Hati. *The Indonesian Journal of Computer Science*, *12*(4), 2096–2106. https://doi.org/10.33022/ijcs.v12i4.3274

Pardede, J., & Ibrahim, R. G. (2020). Implementasi Long Short-Term Memory untuk Identifikasi Berita Hoax Berbahasa Inggris pada Media Sosial. *Journal of Computer Science and Informatics Engineering (J-Cosine)*, *4*(2), 179–187. https://doi.org/10.29303/jcosine.v4i2.361

Purnajaya, A. R., & Pernando, Y. (2023). Analisa Sentimen Informasi Hoaks Pasca Pandemi Covid-19 dengan Text Mining. *Journal of Computer System and Informatics (JoSYC)*, *4*(3), 460–469. https://doi.org/10.47065/josyc.v4i3.3358

Putri, N. F., Vionia, E., & Michael, T. (2020). Pentingnya Kesadaran Hukum Dan Peran Masyarakat Indonesia Dalam Menghadapi Penyebaran Berita Hoax Covid-19. *Media Keadilan: Jurnal Ilmu Hukum*, *11*(1), 98–111. https://doi.org/10.31764/jmk.v11i1.2262

Ramadhan, N. G., Adhinata, F. D., Segara, A. J. T., & Rakhmadani, D. P. (2022). Deteksi Berita Palsu Menggunakan Metode Random Forest dan Logistic Regression. *JURIKOM (Jurnal Riset Komputer)*, *9*(2), 251–256. https://doi.org/10.30865/jurikom.v9i2.3979

Ropikoh, I. A., Abdulhakim, R., Enri, U., & Sulistiyowati, N. (2021). Penerapan Algoritma Support Vector Machine (SVM) untuk Klasifikasi Berita Hoax Covid-19. *Journal of Applied Informatics and Computing*, *5*(1), 64–73. https://doi.org/10.30871/jaic.v5i1.3167

Roy, J., & Junaidi, A. (2020). Pengaruh Terpaan Media Berita Hoax di Instagram terhadap Opini Masyarakat Milenials Akan Sumber Berita. *Koneksi*, *4*(2), 280–285. https://doi.org/10.24912/kn.v4i2.8138

Ruise, A. P., Mashuri, A. S., Sulaiman, M., & Rahman, F. (2023). Studi Komparasi Metode Svm, Logistic Regresion Dan Random Forest Clasifier Untuk Mengklasifikasi Fake News di Twitter. *J I M P - Jurnal Informatika Merdeka Pasuruan*, *7*(2), 64–67. https://doi.org/10.51213/jimp.v7i2.472

Sani, R. R., Pratiwi, Y. A., Winarno, S., Udayanti, E. D., & Alzami, F. (2022). Analisis Perbandingan Algoritma Naive Bayes Classifier dan Support Vector Machine untuk Klasifikasi Berita Hoax pada Berita Online Indonesia. *Jurnal Masyarakat Informatika*, *13*(2), 85–98. https://doi.org/10.14710/jmasif.13.2.47983

Tanggraeni, A. I., & Sitokdana, M. N. N. (2022). Analisis Sentimen Aplikasi E-Government pada Google Play Menggunakan Algoritma Naïve Bayes. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *9*(2), 785–795. https://doi.org/10.35957/jatisi.v9i2.1835

Zhafira, D. F., Rahayudi, B., & Indriati, I. (2021). Analisis Sentimen Kebijakan Kampus Merdeka Menggunakan Naive Bayes dan Pembobotan TF-IDF Berdasarkan Komentar pada Youtube. *Jurnal Sistem Informasi, Teknologi Informasi, Dan Edukasi Sistem Informasi*, *2*(1), 55–63. https://doi.org/10.25126/justsi.v2i1.24