Jurnal Pendidikan dan Teknologi Indonesia (JPTI)
Vol. 4, No. 8, Agustus 2024, Hal. 381-392
DOI: https://doi.org/10.52436/1.jpti.602
p-ISSN: 2775-4227
e-ISSN: 2775-4219

Optimasi Logistic Regression dan Random Forest untuk Deteksi Berita Hoax Berbasis TF-IDF

Arif Mu'amar Wahid[*1], Turino[2], Khabib Adi Nugroho[3], Titi Safitri[4], Darmono[5], Fandy Setyo Utomo[6]

[1,2,3,4,5,6]Magister Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto, Indonesia
Email: [1]arif@amikompurwokerto.ac.id, [2]23MA41D002@students.amikompurwokerto.ac.id,
[3]23MA41D018@students.amikompurwokerto.ac.id,
[4]23MA41D010@students.amikompurwokerto.ac.id,
[5]23MA41D027@students.amikompurwokerto.ac.id, [6]fandy_setyo_utomo@amikompurwokerto.ac.id

Abstrak

Penyebaran berita hoax di era digital menjadi tantangan serius yang memerlukan solusi berbasis teknologi untuk mengidentifikasi dan meminimalkan dampaknya. Penelitian ini bertujuan untuk mengevaluasi performa Logistic Regression (LR) dan Random Forest (RF) dalam mendeteksi berita hoax menggunakan representasi teks berbasis Term Frequency-Inverse Document Frequency (TF-IDF). Hyperparameter tuning diterapkan pada kedua algoritma untuk meningkatkan akurasi, precision, recall, dan F1-score. Dataset yang digunakan terdiri dari berita hoax dan valid dalam bahasa Indonesia, yang telah melalui tahapan preprocessing, termasuk pembersihan teks, penghapusan stopwords, dan stemming. Hasil evaluasi menunjukkan bahwa Logistic Regression, setelah tuning, mencapai akurasi sebesar 95.20%, precision 95.71%, recall 94.48%, dan F1-score 95.09%. Random Forest menunjukkan akurasi sebesar 92.39%, precision 94.39%, recall 89.87%, dan F1-score 92.08%. Logistic Regression unggul dalam keseimbangan antara precision dan recall, sementara Random Forest menunjukkan kekuatan pada precision dengan kemampuan menangani pola data yang lebih kompleks. Teknik TF-IDF terbukti efektif dalam memberikan bobot pada kata-kata yang relevan, membantu algoritma klasifikasi dalam mengenali pola dalam data teks. Penelitian ini juga memiliki dampak praktis dalam memberikan fondasi bagi pengembangan sistem deteksi hoax yang dapat digunakan di aplikasi berbasis NLP, baik untuk kebutuhan akademis maupun implementasi di industri. Penelitian ini berkontribusi pada pengembangan sistem deteksi hoax berbasis Natural Language Processing (NLP), khususnya untuk bahasa Indonesia. Untuk pengembangan lebih lanjut, disarankan memperluas dataset dengan sumber berita yang lebih beragam dan mengeksplorasi algoritma berbasis deep learning seperti LSTM atau Transformer. Secara ilmiah, penelitian ini memberikan kontribusi penting dengan menguji efektivitas hyperparameter tuning dalam meningkatkan akurasi model deteksi hoax. Hasil penelitian ini diharapkan dapat menjadi acuan dalam membangun sistem deteksi hoax yang lebih akurat dan andal.

Kata kunci: *deteksi hoax, hyperparameter tuning, logistic regression, nlp, random forest, tf-idf*

Hyperparameter Optimization of Logistic Regression and Random Forest for Hoax News Detection Using TF-IDF Text Representation

Abstract

The spread of hoaxes in the digital era poses a significant challenge that necessitates technology-based solutions to identify and mitigate its impact. This study aims to evaluate the performance of Logistic Regression (LR) and Random Forest (RF) in detecting hoaxes using Term Frequency-Inverse Document Frequency (TF-IDF) for text representation. Hyperparameter tuning was applied to both algorithms to enhance accuracy, precision, recall, and F1-score. The dataset comprised hoax and valid news articles in Indonesian, processed through text cleaning, stopword removal, and stemming. The evaluation results demonstrate that Logistic Regression, after tuning, achieved an accuracy of 95.20%, precision of 95.71%, recall of 94.48%, and an F1-score of 95.09%. Random Forest achieved an accuracy of 92.39%, precision of 94.39%, recall of 89.87%, and an F1-score of 92.08%. Logistic Regression excelled in balancing precision and recall, while Random Forest showcased strength in precision and its ability to handle more complex data patterns. The TF-IDF technique proved effective in assigning weight to relevant words, aiding classification algorithms in identifying patterns in textual data. This research also has a practical impact in providing a foundation for the development of hoax detection systems that can be used in NLP-based applications, both for academic needs and industrial implementation.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

This research contributes to the development of Natural Language Processing (NLP)-based hoax detection systems, especially for the Indonesian language. For further development, it is recommended to expand the dataset with more diverse news sources and explore deep learning-based algorithms such as LSTM or Transformer. Scientifically, this research makes an important contribution by testing the effectiveness of hyperparameter tuning in improving the accuracy of hoax detection models. The results of this research are expected to be a reference in building a more accurate and reliable hoax detection system.

Keywords: *hoax detection, hyperparameter tuning, logistic regression, nlp, random forest, tf-idf*

1. PENDAHULUAN

Penyebaran hoax atau berita palsu di Indonesia menjadi salah satu tantangan besar di era digital, seiring dengan meningkatnya penggunaan media sosial dan platform digital. Dengan lebih dari 170 juta pengguna internet di Indonesia, media sosial telah menjadi saluran utama penyebaran informasi, termasuk berita palsu yang sering kali digunakan untuk memanipulasi opini publik, menciptakan ketidakstabilan sosial, atau memperoleh keuntungan finansial [1]. Fenomena ini semakin terlihat selama pandemi COVID-19, di mana hoax terkait vaksinasi dan virus menyebar luas, menyebabkan kebingungan serta ketidakpastian di kalangan masyarakat [2]. Statistik menunjukkan bahwa penyebaran hoax di Indonesia sangat signifikan. Survei Masyarakat Telematika Indonesia (Mastel) pada tahun 2017 mencatat bahwa 92,40% hoax tersebar melalui media sosial [3]. Selama pandemi, jumlah pengguna media sosial melonjak, mencapai lebih dari 160 juta pada Januari 2020, yang memperparah paparan masyarakat terhadap informasi palsu [4]. Kondisi ini diperburuk oleh rendahnya literasi digital masyarakat Indonesia, membuat mereka lebih rentan terhadap hoax [5].

Dampak hoax sangat luas, mencakup politik, kesehatan, hingga psikologis. Dalam politik, hoax sering digunakan untuk mempengaruhi persepsi publik terhadap kandidat selama pemilu, sehingga merusak proses demokrasi [6]. Sementara itu, dalam konteks kesehatan, hoax tentang vaksinasi menyebabkan penurunan partisipasi masyarakat dalam program vaksinasi, yang berdampak buruk pada upaya penanganan pandemi COVID-19 [7]. Efek psikologis juga tak terelakkan, seperti meningkatnya kecemasan di masyarakat akibat paparan informasi yang salah. Dalam menghadapi tantangan ini, pendekatan berbasis teknologi untuk deteksi hoax menjadi sangat relevan. Perkembangan Natural Language Processing (NLP) dan Machine Learning (ML) memungkinkan implementasi sistem deteksi otomatis dengan akurasi tinggi [8]. Salah satu pendekatan yang banyak digunakan adalah representasi teks berbasis Term Frequency-Inverse Document Frequency (TF-IDF), yang secara efektif mengidentifikasi pola-pola unik dalam teks berita palsu [9]. TF-IDF menjadi komponen penting dalam proses deteksi hoax karena kemampuannya untuk memberikan bobot lebih pada kata-kata yang unik dan jarang muncul, namun relevan, dalam dokumen tertentu [10]. Ketika dikombinasikan dengan algoritma klasifikasi seperti Logistic Regression atau Random Forest, TF-IDF menunjukkan performa yang signifikan dalam meningkatkan akurasi deteksi [11]. Dengan teknik ini, proses deteksi dapat dilakukan lebih efisien dan dalam skala besar, mengingat tingginya volume informasi digital yang tersebar setiap hari.

Dalam deteksi hoax, pemilihan teknik representasi teks dan algoritma klasifikasi sangat memengaruhi akurasi dan efisiensi model. Salah satu teknik representasi teks yang banyak digunakan adalah Term Frequency-Inverse Document Frequency (TF-IDF), yang memberikan bobot lebih pada kata-kata unik dalam sebuah dokumen. TF-IDF dianggap efektif untuk mengidentifikasi pola dalam berita hoax, namun efektivitasnya dapat bervariasi tergantung pada algoritma klasifikasi yang digunakan [9]. Oleh karena itu, perlu dilakukan evaluasi terhadap kinerja TF-IDF saat digunakan bersama algoritma klasifikasi seperti Logistic Regression dan Random Forest. Logistic Regression adalah algoritma yang sederhana namun efektif untuk klasifikasi. Dengan pendekatan berbasis probabilitas, Logistic Regression memberikan hasil yang mudah diinterpretasikan, menjadikannya ideal untuk analisis awal [8]. Selain itu, kecepatan pelatihannya menjadikannya pilihan efisien, terutama untuk dataset besar. Namun, kelemahan utama Logistic Regression adalah ketidakmampuannya menangani hubungan non-linear antara fitur dan label, yang sering kali ditemukan dalam berita hoax. Dalam situasi ini, Logistic Regression dapat memberikan hasil yang kurang akurat atau bahkan mengalami overfitting jika tidak diatur dengan baik. Berbeda dengan Logistic Regression, Random Forest mampu menangani hubungan non-linear secara lebih baik. Algoritma ini menggabungkan beberapa pohon keputusan untuk memberikan hasil yang lebih stabil dan tahan terhadap overfitting [11]. Selain itu, Random Forest dapat menangani data numerik maupun kategorikal tanpa memerlukan normalisasi, sehingga fleksibel untuk berbagai jenis data. Meski demikian, interpretabilitas Random Forest lebih rendah dibandingkan Logistic Regression, dan waktu pelatihannya lebih lambat, terutama untuk dataset yang sangat besar. Namun, kemampuannya dalam menangkap kompleksitas data menjadikannya pilihan kuat untuk mendeteksi berita hoax.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Studi Karo menunjukkan bahwa kombinasi TF-IDF dengan Bernoulli Naive Bayes meningkatkan akurasi deteksi hoax hingga 16% dibandingkan baseline [12], sementara penelitian Holla memperkenalkan model hibrida berbasis TF-IDF dan AdaBoost yang berhasil mencapai akurasi lebih tinggi dibandingkan metode tradisional [13]. Kedua temuan ini menggarisbawahi fleksibilitas TF-IDF dalam mendukung berbagai kerangka algoritma pembelajaran mesin. Selain itu, TF-IDF juga terbukti efektif dalam integrasinya dengan teknik deep learning. Putra dan Setiawan melaporkan bahwa penggunaan TF-IDF dalam model Long Short-Term Memory (LSTM) dan Gate Recurrent Unit (GRU) menghasilkan akurasi masing-masing sebesar 97,33% dan 96,75% [14], sedangkan penelitian Aji menunjukkan bahwa TF-IDF lebih unggul dibandingkan teknik vektorisasi lainnya, seperti GloVe, untuk tugas deteksi hoax [15]. Studi Tama memvalidasi adaptabilitas TF-IDF dengan kerangka Convolutional Neural Network (CNN) untuk mendeteksi berita hoax di Twitter [16], sementara Firmanesha menyoroti efektivitas TF-IDF dalam meningkatkan performa klasifikasi menggunakan Support Vector Machines (SVM) dan Stochastic Gradient Descent (SGD) [17]. Secara khusus, pada konteks berita hoax di Indonesia, Prayoga et al. melaporkan akurasi hingga 98,5% menggunakan TF-IDF dengan Bernoulli Naive Bayes [18], mempertegas keandalan TF-IDF untuk aplikasi lokal dan global.

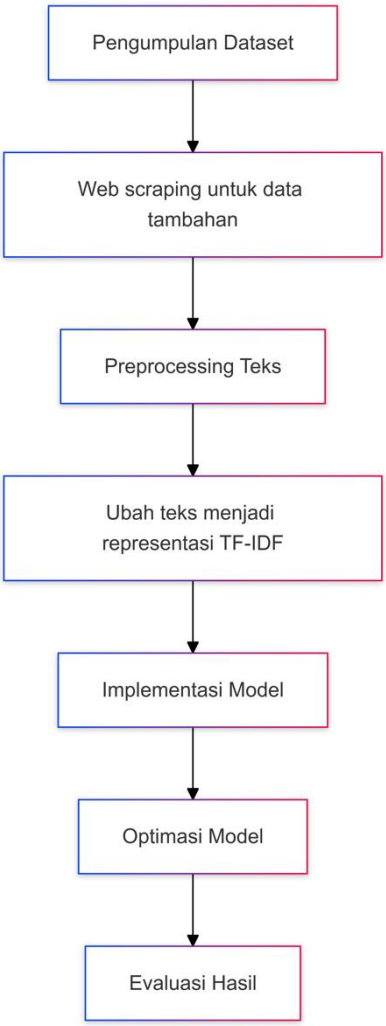
Penelitian ini terinspirasi oleh berbagai studi yang mengaplikasikan metode data mining, machine learning, dan teknik analisis data dalam berbagai konteks, yang memberikan landasan teoretis dan metodologis bagi pengembangan sistem deteksi hoax berbasis NLP. Penelitian [19] menunjukkan efektivitas algoritma Random Forest dan Gradient Boosting dalam memprediksi kerusakan bangunan pasca-gempa, menggarisbawahi kekuatan model ensemble dalam menangani data kompleks. Penelitian [20] menyoroti pentingnya manajemen data terstruktur melalui desain data warehouse, yang relevan dalam mendukung pelaporan akreditasi dan kolaborasi di institusi pendidikan tinggi. Dalam konteks pengolahan citra, penelitian [21] membandingkan kinerja AlexNet dan Inception V3 untuk ekstraksi fitur visual, memberikan wawasan tentang optimalisasi model untuk analisis data non-tekstual. Penelitian [22] menggunakan algoritma seperti Random Forest, SVM, dan neural networks untuk menganalisis faktor yang memengaruhi kepuasan mahasiswa, menyoroti pentingnya pemilihan model yang sesuai untuk pengambilan keputusan berbasis data. Selanjutnya, penelitian [23] mengeksplorasi prediksi tren harga Bitcoin menggunakan ARIMA dan LSTM, menunjukkan potensi kombinasi model statistik dan deep learning untuk analisis data waktu. Studi lain oleh [24] menganalisis dampak strategi diskon pada penilaian konsumen melalui ulasan produk di Amazon, memberikan kontribusi penting dalam memahami perilaku konsumen melalui data besar. Penelitian-penelitian ini mempertegas relevansi teknik machine learning dan analisis data dalam berbagai domain, menjadi inspirasi utama untuk mengembangkan model deteksi hoax yang lebih akurat dan efisien.

Penelitian ini bertujuan untuk mengevaluasi efektivitas Term Frequency-Inverse Document Frequency (TF-IDF) sebagai teknik representasi teks dalam mendeteksi hoax, dengan fokus pada konteks dataset berbahasa Indonesia yang memiliki pola bahasa unik dan belum banyak dijelajahi. Kebaruan penelitian ini terletak pada eksplorasi TF-IDF untuk mendeteksi pola-pola linguistik spesifik dalam berita hoax berbahasa Indonesia. Selain itu, penelitian ini mengintegrasikan TF-IDF dengan dua algoritma klasifikasi, yaitu Logistic Regression dan Random Forest, dan membandingkan performanya secara mendalam. Kebaruan lain adalah pendekatan sistematis untuk memahami peran TF-IDF dalam meningkatkan efektivitas hyperparameter tuning pada kedua algoritma, yang jarang dibahas dalam konteks deteksi hoax. Dengan mengeksplorasi kombinasi TF-IDF dan kedua algoritma tersebut, penelitian ini tidak hanya berkontribusi pada pengembangan model deteksi hoax yang lebih akurat tetapi juga memberikan wawasan baru tentang bagaimana teknik representasi teks dapat dioptimalkan dalam sistem berbasis Natural Language Processing (NLP). Selain itu, penelitian ini memberikan rekomendasi praktis untuk deteksi hoax yang lebih efisien pada dataset berbahasa Indonesia, yang relevansinya semakin meningkat di tengah lonjakan informasi digital lokal.

2. METODE PENELITIAN

Untuk memperjelas metode penelitian, Gambar 1 di bawah ini menggambarkan alur penelitian untuk mendeteksi berita hoax, dimulai dari pengumpulan dataset, dilanjutkan dengan scraping data tambahan, preprocessing teks, transformasi teks menjadi representasi TF-IDF, implementasi model klasifikasi, optimasi model, hingga evaluasi hasil untuk menentukan performa terbaik.

p-ISSN: 2775-4227
e-ISSN: 2775-4219



Gambar 1. Diagram Alir Metode Penelitian

2.1. Pengumpulan Dataset

Tahap awal dalam penelitian ini adalah pengumpulan dataset yang digunakan untuk mendeteksi berita hoax. Penelitian ini menggunakan dua dataset berita berbahasa Indonesia, yaitu 500_berita_Indonesia.csv dan 600_news_with_valid_hoax_label.csv. Kedua dataset ini dipilih karena keduanya mengandung data berita dengan label yang jelas sebagai valid atau hoax, sehingga relevan untuk tujuan penelitian ini. Namun, karena dataset berasal dari sumber yang berbeda, langkah-langkah khusus diperlukan untuk memastikan keduanya dapat digabungkan secara konsisten. Langkah pertama adalah memuat dataset menggunakan bahasa pemrograman Python. Dataset pertama, 500_berita_Indonesia.csv, dimuat menggunakan encoding utf-8-sig dengan delimiter ;, karena format CSV yang digunakan membutuhkan penanganan khusus untuk membaca data dengan benar. Dataset kedua, 600_news_with_valid_hoax_label.csv, dimuat menggunakan encoding latin1 untuk menangani perbedaan format encoding. Dalam kedua kasus, opsi on_bad_lines='skip' digunakan untuk melewati baris yang tidak sesuai dengan format, sehingga memastikan data yang dimuat bersih dan bebas dari kesalahan struktur.

Setelah kedua dataset berhasil dimuat, langkah berikutnya adalah memverifikasi struktur data dan menyusun ulang kolom agar konsisten. Pada dataset kedua, kolom kategori dan berita ditukar posisinya agar sesuai dengan urutan kolom pada dataset pertama. Langkah ini penting untuk memastikan proses penggabungan kedua dataset dapat dilakukan tanpa kendala. Proses penggabungan dilakukan menggunakan fungsi concat() dari library Pandas, yang memungkinkan penggabungan data secara vertikal. Hasil penggabungan ini menghasilkan satu dataset besar yang mencakup semua data dari kedua sumber, dengan total jumlah baris yang meningkat secara signifikan. Dataset gabungan kemudian disimpan dalam file CSV baru dengan nama combined_berita_dataset.csv menggunakan encoding utf-8. Penyimpanan ini penting agar dataset yang telah diproses dapat digunakan kembali tanpa perlu mengulang tahapan penggabungan. Untuk memastikan bahwa

p-ISSN: 2775-4227
e-ISSN: 2775-4219

penggabungan telah berhasil, dilakukan pemeriksaan ukuran dan struktur dataset, termasuk jumlah baris dan kolom. Selain itu, tipe data dari setiap kolom diperiksa untuk memastikan formatnya sesuai dengan kebutuhan analisis selanjutnya. Hasil dari tahapan ini adalah dataset yang siap untuk diproses lebih lanjut, mencakup berita valid dan hoax dengan struktur data yang seragam dan konsisten.

Selanjutnya adalah proses pengumpulan data tambahan dengan cara scraping dari website turnbackhoax.id. Proses ini dilakukan untuk mendapatkan data berita hoax yang belum ada di dataset awal. Setiap langkah kode berikut dijelaskan untuk memberikan pemahaman rinci mengenai proses pengumpulan data tambahan ini. Proses dimulai dengan inisialisasi pustaka-pustaka yang dibutuhkan, seperti requests untuk mengirim permintaan HTTP, BeautifulSoup untuk mengekstrak konten dari HTML, dan pandas untuk pengolahan data. Selain itu, modul Retry digunakan untuk menangani pengulangan permintaan jika terjadi kegagalan jaringan. URL utama situs web ditentukan dengan variabel base_url, sementara parameter paginasi seperti start_page dan page_limit digunakan untuk menentukan rentang halaman yang akan di-scrape. Untuk menyimpan data yang dikumpulkan, dibuat daftar kosong seperti links, titles, dates, authors, dan sebagainya, yang nantinya akan diisi dengan data dari artikel yang diambil.

Dalam hal ini, konfigurasi session dengan pengaturan retry sangat penting untuk menghindari kesalahan selama scraping. Dengan Retry yang diatur untuk mencoba ulang hingga 5 kali, permintaan dapat tetap berjalan meskipun server mengalami masalah sementara. Fungsi khusus get_full_content_and_hoax digunakan untuk mengambil konten lengkap dari setiap artikel dan mendeteksi klaim hoax melalui berbagai metode, termasuk pencarian elemen <blockquote> atau teks di antara kata-kata kunci seperti "Narasi" dan "Penjelasan". Setiap artikel yang ditemukan diproses untuk mengumpulkan informasi seperti tautan, judul, tanggal, dan penulis. Penundaan waktu sebesar 1 detik antara permintaan digunakan untuk mencegah server mendeteksi aktivitas scraping sebagai bot yang berlebihan. Setelah semua data terkumpul, data disusun dalam bentuk DataFrame menggunakan pustaka pandas, dan disimpan ke dalam file CSV dengan format tertentu. Untuk memastikan tidak ada data yang terlewat, metode tambahan diterapkan seperti pencarian pola dengan regex untuk mendeteksi konten hoax yang mungkin tidak teridentifikasi sebelumnya. Hasil akhirnya adalah dataset baru yang berisi data dari semua metode yang berhasil. Langkah terakhir adalah penggabungan dataset hasil scraping dengan dataset asli. Dataset hasil scraping dimuat kembali dari file CSV, sementara dataset asli juga diperiksa dan dibersihkan, termasuk penghapusan baris baru dalam kolom berita. Kedua dataset kemudian digabungkan menggunakan fungsi pd.concat, dengan opsi ignore_index=True untuk menciptakan indeks baru yang berurutan. Dataset gabungan ini disimpan ke dalam file CSV baru untuk memastikan bahwa semua data hoax yang relevan tersedia dalam satu tempat, sehingga siap untuk dianalisis lebih lanjut pada tahap berikutnya.

2.2. Preprocessing Teks

Tahap awal dalam preprocessing teks adalah memuat dataset gabungan yang telah disimpan sebelumnya. Dataset ini berisi berita dan kategorinya, yang masing-masing dilabeli sebagai hoax atau valid. Dataset dimuat menggunakan pustaka Pandas, memastikan struktur data tetap konsisten dari tahap penggabungan sebelumnya. Langkah ini dilakukan untuk menghindari kesalahan dalam proses lebih lanjut. Setelah dataset dimuat, kolom yang relevan seperti berita dan kategori akan menjadi fokus utama untuk pemrosesan lebih lanjut. Dataset ini memberikan dasar untuk seluruh tahapan preprocessing berikutnya. Langkah selanjutnya adalah pembersihan teks (text cleaning), yang melibatkan penghapusan simbol, angka, dan tanda baca yang tidak relevan. Proses ini dilakukan dengan fungsi khusus menggunakan ekspresi reguler (regex), yang memastikan hanya huruf alfabet yang tersisa dalam teks. Semua huruf kemudian diubah menjadi huruf kecil untuk menjaga konsistensi, menghindari perbedaan yang disebabkan oleh kapitalisasi. Tahap ini bertujuan untuk mengurangi noise dalam data, sehingga teks menjadi lebih bersih dan siap untuk analisis lebih lanjut. Contohnya, kalimat "Berita HOAX! Cek Fakta: Tidak Benar" akan dibersihkan menjadi "berita hoax cek fakta tidak benar."

Setelah teks dibersihkan, dilakukan penghapusan stopwords menggunakan pustaka Sastrawi. Stopwords seperti "dan," "atau," dan "adalah" dihapus karena tidak memberikan kontribusi informatif terhadap analisis teks. Proses ini membantu mengurangi jumlah fitur yang tidak relevan dan meningkatkan efisiensi algoritma klasifikasi. Selanjutnya, teks yang telah dihapus stopwords-nya diproses menggunakan stemming. Dengan menggunakan fitur stemmer dari Sastrawi, kata-kata seperti "menyelesaikan" diubah menjadi "selesai." Proses ini menyamakan berbagai bentuk kata menjadi bentuk dasarnya, sehingga model dapat mengenali kata-kata dengan makna yang sama meskipun memiliki bentuk yang berbeda. Tahap akhir preprocessing adalah transformasi TF-IDF (Term Frequency-Inverse Document Frequency), yang dilakukan pada teks yang telah melalui proses pembersihan, penghapusan stopwords, dan stemming. TF-IDF digunakan untuk menghitung bobot setiap kata dalam dokumen relatif terhadap kumpulan dokumen. Dengan menerapkan metode fit_transform dari pustaka Scikit-learn, data latih dikonversi menjadi matriks TF-IDF, sementara data uji diubah menggunakan model TF-IDF yang telah dibuat dari data latih. Hasilnya adalah representasi numerik dari teks

p-ISSN: 2775-4227
e-ISSN: 2775-4219

yang siap digunakan sebagai input ke model klasifikasi. Untuk efisiensi penyimpanan, matriks TF-IDF disimpan dalam format .npz yang mendukung struktur matriks sparse. Output dari tahapan ini adalah dataset yang telah direpresentasikan dalam bentuk fitur numerik, siap untuk digunakan dalam pelatihan dan evaluasi model klasifikasi.

2.3. Algoritma Klasifikasi

Pada tahap implementasi algoritma klasifikasi, penelitian ini menggunakan dua model utama, yaitu Logistic Regression dan Random Forest. Kedua algoritma ini dipilih karena memiliki karakteristik yang berbeda, yang memungkinkan perbandingan kinerja pada tugas deteksi hoax. Logistic Regression, yang merupakan model klasifikasi linier, sering digunakan sebagai baseline dalam klasifikasi biner. Model ini mengasumsikan hubungan linier antara fitur input dan log-odds dari output, menjadikannya sederhana dan cepat. Sebagai pembanding, Random Forest adalah model berbasis ensemble yang menggunakan beberapa pohon keputusan untuk meningkatkan akurasi dan stabilitas, sehingga lebih efektif dalam menangani data dengan pola kompleks.

Model Logistic Regression diimplementasikan menggunakan pustaka Scikit-learn. Parameter utama yang digunakan adalah max_iter=200, yang meningkatkan iterasi maksimum model. Hal ini memastikan model memiliki cukup waktu untuk mencapai konvergensi selama proses pelatihan, terutama pada dataset besar yang mungkin memerlukan iterasi tambahan untuk menghasilkan model yang optimal. Data latih yang telah ditransformasi menggunakan TF-IDF digunakan untuk melatih model, sementara data uji digunakan untuk mengevaluasi performa. Fungsi khusus evaluate_model diterapkan untuk melatih model, membuat prediksi, dan menghitung metrik evaluasi seperti akurasi, precision, recall, F1 score, serta confusion matrix. Logistic Regression dipilih untuk mengukur performa dasar dari sistem deteksi hoax karena kemampuannya memberikan hasil yang mudah diinterpretasikan dan efisien dalam waktu pelatihan.

Random Forest diimplementasikan dengan menggunakan parameter n_estimators=100, yang menentukan jumlah pohon keputusan dalam model. Semakin banyak pohon, semakin stabil hasil prediksi, namun juga meningkatkan waktu pelatihan. Parameter random_state=42 digunakan untuk memastikan hasil yang konsisten di setiap eksekusi. Data latih dan data uji diproses menggunakan pendekatan yang sama seperti pada Logistic Regression, dengan fungsi evaluate_model digunakan untuk mengevaluasi performa. Random Forest memiliki keunggulan dalam menangani data yang kompleks dan hubungan non-linear, menjadikannya pilihan yang baik untuk mendeteksi pola dalam dataset hoax yang beragam. Selain itu, model ini dapat memberikan informasi penting tentang fitur mana yang paling berkontribusi terhadap hasil prediksi.

Kedua model dievaluasi dengan menggunakan metrik seperti akurasi, precision, recall, dan F1 score, untuk memberikan gambaran lengkap tentang kinerja model. Logistic Regression memberikan hasil yang cepat dan mudah diinterpretasikan, namun cenderung kurang efektif pada data dengan hubungan non-linear. Sebaliknya, Random Forest memberikan hasil yang lebih stabil dan seringkali lebih akurat, meskipun waktu pelatihannya lebih lama. Dengan perbandingan ini, penelitian dapat menentukan algoritma mana yang lebih efektif untuk mendeteksi hoax pada dataset yang telah direpresentasikan menggunakan TF-IDF. Kombinasi dari pendekatan ini memberikan wawasan yang lebih mendalam tentang efektivitas representasi teks dan algoritma klasifikasi pada tugas deteksi hoax.

2.4. Evaluasi Model

Evaluasi model dimulai dengan pembagian dataset menjadi data pelatihan dan data pengujian. Proses ini dilakukan menggunakan fungsi train_test_split dari pustaka Scikit-learn, dengan parameter test_size=0.2, yang berarti 80% data digunakan untuk pelatihan, sementara 20% sisanya digunakan untuk pengujian. Pembagian data ini penting untuk memastikan bahwa model memiliki data yang cukup untuk belajar (pelatihan) sekaligus data yang tidak terlihat sebelumnya untuk mengevaluasi kinerjanya (pengujian). Parameter tambahan seperti random_state=42 digunakan untuk menjaga konsistensi dalam proses pembagian data, sehingga hasil evaluasi model tetap dapat direproduksi. Setelah pembagian dataset, metrik evaluasi yang digunakan mencakup akurasi, precision, recall, dan F1-score. Akurasi menghitung persentase prediksi yang benar terhadap total data uji, memberikan gambaran umum tentang performa model. Namun, akurasi saja sering kali tidak cukup, terutama jika dataset tidak seimbang antara label hoax dan valid. Precision digunakan untuk mengukur proporsi prediksi hoax yang benar-benar hoax, memberikan fokus pada prediksi positif yang relevan. Recall, di sisi lain, menghitung seberapa baik model mampu mendeteksi seluruh data hoax yang sebenarnya, memastikan bahwa model tidak melewatkan banyak kasus hoax.

Metrik F1-score digunakan untuk menyeimbangkan precision dan recall, terutama dalam kasus di mana ada ketidakseimbangan antara data hoax dan valid. F1-score adalah rata-rata harmonis dari precision dan recall, memberikan penilaian yang lebih komprehensif terhadap kemampuan model dalam mendeteksi hoax. Selain itu, confusion matrix juga disertakan untuk memberikan visualisasi distribusi prediksi model, termasuk jumlah true

p-ISSN: 2775-4227
e-ISSN: 2775-4219

positive (TP), false positive (FP), true negative (TN), dan false negative (FN). Confusion matrix membantu mengidentifikasi kesalahan model, misalnya apakah model lebih sering salah memprediksi hoax sebagai valid atau sebaliknya. Evaluasi dilakukan pada kedua algoritma, Logistic Regression dan Random Forest, menggunakan fungsi evaluasi yang sama untuk memastikan perbandingan yang adil. Data pelatihan digunakan untuk melatih model pada representasi teks TF-IDF, sementara data pengujian digunakan untuk mengevaluasi performa berdasarkan metrik yang disebutkan. Hasil evaluasi ini memberikan pemahaman mendalam tentang kelebihan dan kekurangan masing-masing algoritma dalam mendeteksi hoax, serta memberikan dasar untuk menentukan algoritma mana yang lebih cocok digunakan pada skenario dunia nyata. Dengan pendekatan evaluasi yang terstruktur, penelitian ini memastikan bahwa hasilnya dapat dipercaya dan relevan untuk tujuan deteksi hoax.

2.5. Optimasi dengan Hyperparameter Tuning

Optimasi hyperparameter dilakukan untuk meningkatkan performa model klasifikasi Logistic Regression (LR) dan Random Forest (RF) yang diterapkan pada dataset dengan representasi teks berbasis TF-IDF. Langkah ini bertujuan untuk menemukan kombinasi parameter terbaik yang dapat meningkatkan akurasi, precision, recall, dan F1-score dalam mendeteksi berita hoax. Pada Logistic Regression, tuning dilakukan pada parameter utama seperti C (Regularization Strength), solver, dan penalty. Parameter C mengontrol kekuatan regulasi dalam model, di mana nilai kecil menunjukkan regulasi yang lebih kuat untuk menghindari overfitting, sementara nilai besar mengurangi regulasi. Rentang nilai yang diuji adalah [0.01,0.1,1,10,100] untuk menemukan keseimbangan antara bias dan varians. Solver yang diuji termasuk lbfgs dan liblinear, dipilih berdasarkan kompatibilitas dengan penalty L2 serta efisiensi dalam menangani dataset yang berbeda ukuran. Solver lbfgs cocok untuk dataset besar, sedangkan liblinear lebih efisien pada dataset kecil hingga menengah. Penalty L2, yang memberikan regularisasi Ridge, digunakan untuk meningkatkan stabilitas model, terutama pada dataset hasil TF-IDF yang memiliki dimensi tinggi. L2 membantu mencegah nilai koefisien menjadi terlalu besar, yang sering kali terjadi pada data berdimensi besar.

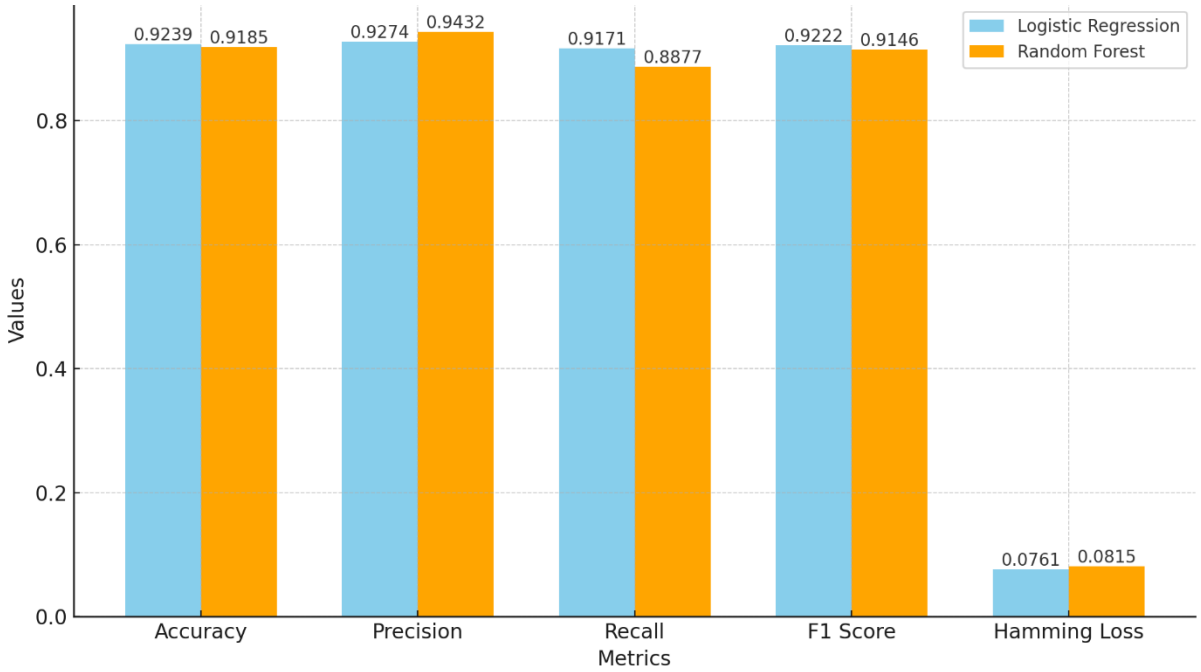
Pada Random Forest, parameter yang disesuaikan meliputi n_estimators, max_depth, dan min_samples_split. n_estimators menentukan jumlah pohon keputusan dalam model, dengan nilai yang diuji pada rentang [50,100,150,200,300]. Semakin banyak pohon memberikan prediksi yang lebih stabil, meskipun memerlukan waktu komputasi lebih lama. Max_depth membatasi kedalaman maksimum pohon untuk mencegah overfitting. Nilai yang diuji mencakup [10,20,30,None], dengan None mengizinkan pohon tumbuh hingga sempurna. Pembatasan kedalaman berguna untuk menangkap pola data yang penting tanpa membuat model terlalu kompleks. Min_samples_split mengatur jumlah minimum sampel yang diperlukan untuk membagi simpul, dengan nilai yang diuji pada rentang [2,5,10]. Parameter ini mengurangi risiko overfitting dengan memastikan bahwa simpul tidak terbagi jika data terlalu sedikit, sehingga menjaga generalisasi model. Optimasi dilakukan menggunakan grid search untuk menguji kombinasi parameter secara sistematis dan memilih kombinasi yang memberikan performa terbaik berdasarkan metrik evaluasi. Pendekatan ini memastikan bahwa model tidak hanya akurat, tetapi juga efisien dan mampu menangani variasi dalam data secara efektif.

3. HASIL DAN PEMBAHASAN

3.1. Hasil Evaluasi Sebelum Optimasi

Pada tahap evaluasi awal, performa kedua model klasifikasi, Logistic Regression (LR) dan Random Forest (RF), dianalisis berdasarkan representasi teks menggunakan TF-IDF. Evaluasi dilakukan menggunakan metrik utama seperti akurasi, precision, recall, F1-score, dan Hamming loss. Selain itu, analisis lebih rinci dilakukan dengan menggunakan matriks kebingungan (confusion matrix) untuk memahami distribusi prediksi benar dan salah pada kelas positif (hoax) dan negatif (valid). Hasil evaluasi metrik sebelum optimasi ditunjukkan pada Gambar 2. Logistic Regression menunjukkan akurasi sebesar 92.39%, dengan precision sebesar 92.74% dan recall sebesar 91.71%. F1-score yang dihasilkan adalah 92.22%, mencerminkan keseimbangan antara precision dan recall. Hamming loss, yang mengukur proporsi kesalahan prediksi, adalah 7.61%, menunjukkan bahwa model ini memiliki tingkat kesalahan yang relatif rendah. Berdasarkan matriks kebingungan, model ini berhasil memprediksi 522 true positives (TP) dan 498 true negatives (TN), namun masih terdapat 39 false positives (FP) dan 45 false negatives (FN).

p-ISSN: 2775-4227
e-ISSN: 2775-4219



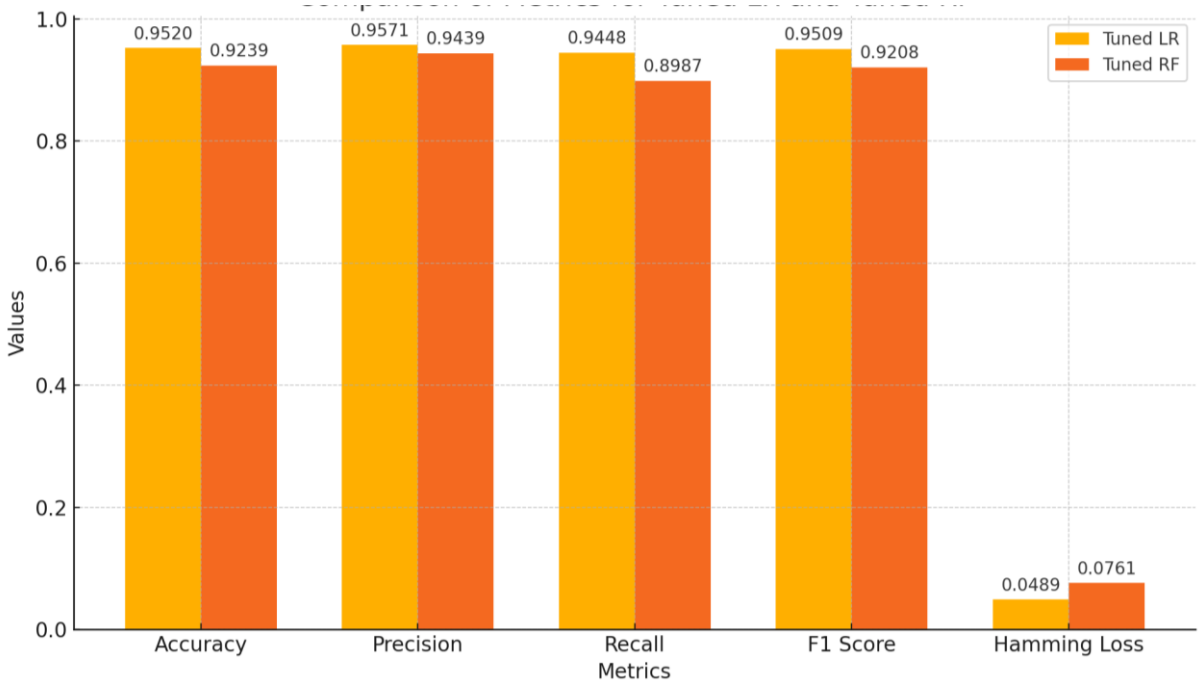
Gambar 2. Hasil Evaluasi Metrik Sebelum Optimasi

Di sisi lain, Random Forest memberikan akurasi yang sedikit lebih rendah, yaitu 91.85%, namun memiliki precision yang lebih tinggi dibandingkan Logistic Regression, yaitu 94.32%. Ini menunjukkan bahwa Random Forest lebih baik dalam memprediksi hoax dengan benar dibandingkan Logistic Regression. Namun, recall yang dihasilkan lebih rendah, yaitu 88.77%, mengindikasikan bahwa model ini cenderung melewatkan beberapa kasus hoax. F1-score model ini adalah 91.46%, sedikit di bawah Logistic Regression. Hamming loss sebesar 8.15% menunjukkan tingkat kesalahan yang sedikit lebih tinggi dibandingkan Logistic Regression. Berdasarkan matriks kebingungan, Random Forest memprediksi 532 TP dan 482 TN, namun memiliki lebih banyak FN (61) meskipun FP lebih sedikit (29) dibandingkan Logistic Regression. Secara keseluruhan, kedua model menunjukkan performa yang baik dengan akurasi lebih dari 91%. Logistic Regression memiliki keseimbangan yang lebih baik antara precision dan recall, sementara Random Forest unggul dalam precision namun memiliki recall yang lebih rendah. Perbedaan ini mencerminkan karakteristik model, di mana Logistic Regression lebih sederhana dan cepat dalam pelatihan, sementara Random Forest mampu menangkap kompleksitas data dengan lebih baik tetapi dengan risiko melewatkan beberapa data penting. Evaluasi ini memberikan dasar untuk mengoptimalkan kedua model pada tahap selanjutnya, dengan fokus pada pengurangan false negatives untuk meningkatkan recall tanpa mengorbankan precision.

3.2. Hasil Evaluasi Setelah Optimasi Hyperparameter Tuning

Setelah dilakukan hyperparameter tuning, performa kedua model klasifikasi, Logistic Regression (LR) dan Random Forest (RF), menunjukkan peningkatan dalam mendeteksi berita hoax dengan representasi teks berbasis TF-IDF. Evaluasi dilakukan menggunakan metrik utama seperti akurasi, precision, recall, F1-score, dan Hamming loss, serta matriks kebingungan untuk analisis lebih mendalam terhadap distribusi prediksi. Hasil evaluasi metrik setelah optimasi ditunjukkan pada Gambar 3. Pada Logistic Regression yang telah dioptimasi, akurasi meningkat menjadi 95.20%, dibandingkan dengan 92.39% sebelum tuning. Precision juga menunjukkan peningkatan signifikan menjadi 95.71%, menunjukkan bahwa model lebih akurat dalam memprediksi berita hoax tanpa terlalu banyak menghasilkan false positives. Recall model meningkat menjadi 94.48%, menunjukkan kemampuan model yang lebih baik dalam mendeteksi seluruh berita hoax yang sebenarnya. F1-score, yang mencerminkan keseimbangan antara precision dan recall, naik menjadi 95.09%, mengindikasikan performa keseluruhan yang sangat baik. Hamming loss turun menjadi 4.89%, yang berarti proporsi kesalahan prediksi berkurang. Berdasarkan matriks kebingungan, model ini berhasil memprediksi 538 true positives (TP) dan 513 true negatives (TN), dengan jumlah false positives (FP) dan false negatives (FN) yang masing-masing berkurang menjadi 23 dan 30.

p-ISSN: 2775-4227
e-ISSN: 2775-4219



Gambar 3. Hasil Evaluasi Metrik Setelah Optimasi

Sementara itu, Random Forest yang telah dioptimasi menunjukkan peningkatan yang lebih terbatas dibandingkan Logistic Regression. Akurasi model meningkat sedikit menjadi 92.39%, dengan precision mencapai 94.39%, menunjukkan kemampuan model untuk meminimalkan false positives. Namun, recall model hanya sedikit meningkat menjadi 89.87%, yang berarti model masih cenderung melewatkan beberapa kasus hoax. F1-score juga meningkat menjadi 92.08%, menunjukkan keseimbangan performa yang lebih baik dibandingkan sebelum tuning. Hamming loss tetap sama, yaitu 7.61%, menunjukkan bahwa jumlah kesalahan prediksi secara keseluruhan tidak berubah secara signifikan. Berdasarkan matriks kebingungan, Random Forest memprediksi 532 TP dan 488 TN, sementara jumlah FP dan FN masing-masing adalah 29 dan 55. Secara keseluruhan, optimasi hyperparameter memberikan dampak positif yang lebih besar pada Logistic Regression dibandingkan Random Forest. Logistic Regression berhasil meningkatkan akurasi, precision, dan recall secara signifikan, menjadikannya model yang lebih andal dalam mendeteksi berita hoax. Sementara itu, Random Forest tetap menunjukkan performa yang baik, terutama dalam precision, namun recall yang lebih rendah menunjukkan bahwa model ini masih memerlukan perhatian khusus untuk menangkap lebih banyak kasus hoax. Hasil ini menunjukkan bahwa optimasi hyperparameter tuning memainkan peran penting dalam meningkatkan performa model, terutama untuk dataset yang kompleks seperti berita hoax.

3.3. Perbandingan Algoritma

Logistic Regression (LR) dan Random Forest (RF) masing-masing memiliki kelebihan dan kekurangan yang memengaruhi kinerja mereka dalam mendeteksi berita hoax. Logistic Regression merupakan algoritma linier yang sederhana, cepat, dan efisien. Dengan asumsi hubungan linier antara fitur dan output, Logistic Regression unggul dalam situasi di mana hubungan antar fitur mudah diidentifikasi. Setelah dioptimasi, Logistic Regression menunjukkan performa yang sangat baik, terutama dalam precision dan recall, yang menghasilkan F1-score yang lebih tinggi dibandingkan Random Forest. Kecepatan pelatihan model ini juga menjadi keuntungan tersendiri ketika diterapkan pada dataset besar. Namun, kelemahan utama Logistic Regression adalah ketergantungannya pada asumsi hubungan linier. Dalam kasus dataset yang memiliki pola non-linear, model ini cenderung kurang mampu menangkap kompleksitas data, sehingga berisiko melewatkan pola-pola penting. Hal ini dapat terlihat pada recall awal sebelum tuning, di mana Logistic Regression menunjukkan kelemahan dalam mendeteksi seluruh berita hoax. Dengan bantuan hyperparameter tuning, kelemahan ini dapat diminimalkan, meskipun kemampuan untuk menangani data kompleks tetap lebih terbatas dibandingkan algoritma seperti Random Forest.

Di sisi lain, Random Forest memiliki keunggulan dalam menangani data dengan pola non-linear dan kompleksitas tinggi. Algoritma ini menggunakan pendekatan ensemble, di mana beberapa pohon keputusan digabungkan untuk menghasilkan prediksi yang lebih stabil dan akurat. Keunggulan Random Forest terlihat dari

p-ISSN: 2775-4227
e-ISSN: 2775-4219

precision yang konsisten tinggi, baik sebelum maupun setelah tuning, menunjukkan bahwa model ini sangat efektif dalam meminimalkan false positives. Selain itu, Random Forest memiliki kemampuan unik untuk memberikan informasi tentang pentingnya fitur, yang dapat digunakan untuk analisis lebih lanjut terhadap dataset. Namun, Random Forest juga memiliki kelemahan, yaitu waktu pelatihan yang lebih lama dibandingkan Logistic Regression, terutama untuk dataset besar. Selain itu, recall yang lebih rendah menunjukkan bahwa model ini masih cenderung melewatkan beberapa berita hoax. Kompleksitas model juga membuatnya lebih sulit untuk diinterpretasikan dibandingkan Logistic Regression. Dengan demikian, meskipun Random Forest unggul dalam stabilitas dan fleksibilitas, Logistic Regression menawarkan solusi yang lebih cepat dan efisien dengan performa keseluruhan yang lebih baik setelah tuning. Perbandingan ini menunjukkan bahwa pemilihan algoritma sangat bergantung pada kebutuhan spesifik dan karakteristik dataset yang digunakan.

3.4. Dampak Hyperparameter Tuning

Hyperparameter tuning memiliki dampak signifikan dalam meningkatkan kinerja model klasifikasi, baik untuk Logistic Regression (LR) maupun Random Forest (RF). Pada Logistic Regression, tuning parameter seperti C, solver, dan penalty memungkinkan model untuk menyeimbangkan bias dan variance dengan lebih baik. Dengan nilai regulasi yang optimal (C), model dapat menghindari overfitting pada data latih sekaligus mempertahankan generalisasi yang baik pada data uji. Setelah tuning, Logistic Regression menunjukkan peningkatan akurasi hingga 95.20% dan F1-score hingga 95.09%, mencerminkan kemampuan model untuk secara konsisten memberikan prediksi yang akurat pada berita hoax maupun valid. Pada Random Forest, tuning parameter seperti n_estimators, max_depth, dan min_samples_split membantu model mengurangi kompleksitas yang tidak perlu dan meningkatkan stabilitas prediksi. Penyesuaian jumlah pohon keputusan (n_estimators) memberikan keseimbangan antara akurasi dan efisiensi komputasi, sementara pengaturan kedalaman maksimum (max_depth) membantu mengontrol overfitting pada data latih. Meski peningkatan performa pada Random Forest tidak sebesar Logistic Regression, model ini tetap menunjukkan peningkatan precision hingga 94.39%, yang menunjukkan kemampuan model untuk mengurangi false positives setelah tuning. Dampak utama hyperparameter tuning adalah peningkatan efisiensi dan efektivitas model dalam menangani data yang kompleks. Logistic Regression mendapatkan keuntungan signifikan dari tuning, terutama dalam hal recall, yang menunjukkan kemampuan model untuk mendeteksi lebih banyak berita hoax. Sementara itu, Random Forest menunjukkan peningkatan stabilitas prediksi tanpa mengorbankan terlalu banyak akurasi. Hasil ini menunjukkan bahwa tuning hyperparameter tidak hanya penting untuk meningkatkan performa model, tetapi juga untuk menyesuaikan model dengan kebutuhan spesifik tugas klasifikasi, seperti mendeteksi hoax secara akurat pada dataset besar dan kompleks.

3.5. Pengaruh Teknik TF-IDF

Teknik Term Frequency-Inverse Document Frequency (TF-IDF) terbukti memiliki pengaruh yang signifikan dalam meningkatkan hasil klasifikasi, terutama dalam konteks deteksi berita hoax. TF-IDF memberikan bobot lebih pada kata-kata yang jarang muncul namun memiliki relevansi tinggi dalam membedakan antara berita hoax dan valid. Dengan pendekatan ini, kata-kata unik dalam dataset menjadi lebih menonjol, sehingga membantu algoritma klasifikasi seperti Logistic Regression dan Random Forest untuk lebih efektif dalam mengenali pola dalam data teks. Sebagai contoh, kata-kata yang spesifik terhadap hoax dapat diberi bobot lebih tinggi, memungkinkan model untuk mengidentifikasi berita palsu dengan lebih akurat. Efektivitas TF-IDF juga terlihat pada peningkatan hasil evaluasi kedua algoritma yang digunakan. Pada Logistic Regression, TF-IDF mampu menangkap hubungan linier yang signifikan antara kata-kata dengan kategori berita, menghasilkan akurasi tinggi baik sebelum maupun setelah tuning. Hal serupa terjadi pada Random Forest, di mana TF-IDF membantu model menangkap pola non-linear yang kompleks dalam teks. Tanpa representasi yang baik seperti TF-IDF, algoritma ini mungkin kesulitan mengenali struktur data teks yang beragam dan tidak terstruktur. Transformasi data teks menjadi fitur numerik melalui TF-IDF memberikan fondasi yang kuat bagi algoritma untuk belajar dari data, terlepas dari kompleksitas dan ukuran dataset.

3.6. Pembahasan

Hasil penelitian ini menunjukkan bahwa penggunaan TF-IDF sebagai teknik representasi teks memberikan performa klasifikasi yang signifikan, terutama dalam mendeteksi berita hoax menggunakan Logistic Regression (LR) dan Random Forest (RF). Temuan ini sejalan dengan literatur terdahulu, seperti yang dilaporkan oleh Karo [12], yang menunjukkan bahwa kombinasi TF-IDF dengan Bernoulli Naive Bayes meningkatkan akurasi hingga 16% dibandingkan baseline. Holla [13] juga melaporkan keunggulan model hibrida berbasis TF-IDF dan AdaBoost dalam meningkatkan akurasi dibandingkan metode tradisional. Hal ini menegaskan fleksibilitas TF-

p-ISSN: 2775-4227
e-ISSN: 2775-4219

IDF dalam mendukung berbagai kerangka algoritma pembelajaran mesin. Dalam penelitian ini, TF-IDF membantu algoritma menangkap pola-pola relevan pada dataset berita hoax berbahasa Indonesia, yang konsisten dengan efektivitas TF-IDF dalam mengidentifikasi kata-kata signifikan yang sebelumnya dilaporkan pada studi-studi tersebut. Selain itu, hasil penelitian ini juga mendukung literatur yang menunjukkan keunggulan TF-IDF dalam integrasi dengan teknik deep learning. Penelitian [14] menemukan bahwa TF-IDF meningkatkan akurasi model Long Short-Term Memory (LSTM) dan Gate Recurrent Unit (GRU) hingga masing-masing 97,33% dan 96,75%, sementara penelitian [15] menyoroti keunggulan TF-IDF dibandingkan teknik vektorisasi lainnya, seperti GloVe, dalam tugas deteksi hoax. Dalam konteks penggunaan pada algoritma yang lebih tradisional seperti Support Vector Machines (SVM) dan Stochastic Gradient Descent (SGD), penelitian [17] menunjukkan peningkatan performa yang signifikan dengan TF-IDF. Studi [18], yang memfokuskan pada berita hoax di Indonesia, juga memperkuat validitas TF-IDF dengan melaporkan akurasi hingga 98,5% menggunakan Bernoulli Naive Bayes. Hasil penelitian ini menambahkan bukti lebih lanjut tentang adaptabilitas TF-IDF pada dataset lokal, seperti berita hoax berbahasa Indonesia, sekaligus mendukung literatur global tentang efektivitas TF-IDF dalam tugas klasifikasi teks.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa Logistic Regression dan Random Forest memiliki performa yang baik dalam mendeteksi berita hoax dengan representasi teks berbasis TF-IDF. Logistic Regression unggul dalam memberikan keseimbangan antara precision dan recall, dengan hasil evaluasi yang menunjukkan akurasi sebesar 95.20% setelah dilakukan hyperparameter tuning. Penelitian ini menegaskan bahwa hyperparameter tuning dapat secara signifikan meningkatkan performa deteksi hoax berbasis Logistic Regression. Random Forest, meskipun memiliki precision yang sedikit lebih tinggi dibandingkan Logistic Regression, menunjukkan recall yang lebih rendah, yang mengindikasikan model ini cenderung melewatkan beberapa kasus hoax. Temuan ini menegaskan bahwa Logistic Regression, dengan tuning yang tepat, dapat menjadi pilihan utama untuk tugas klasifikasi berita hoax, sementara Random Forest lebih cocok untuk menangani data yang lebih kompleks dengan penekanan pada minimisasi false positives.

Penelitian ini memberikan kontribusi signifikan dalam pengembangan sistem deteksi hoax berbasis NLP, khususnya dalam konteks berita berbahasa Indonesia. Dengan menggunakan TF-IDF sebagai teknik representasi teks, penelitian ini menunjukkan bahwa algoritma sederhana seperti Logistic Regression dapat mencapai performa yang kompetitif, asalkan didukung oleh teknik preprocessing dan tuning yang tepat. Selain itu, penggunaan Random Forest sebagai pembanding memberikan wawasan tambahan tentang cara menangani pola non-linear dalam data. Hasil penelitian ini dapat menjadi dasar bagi pengembang sistem deteksi hoax untuk memilih algoritma yang sesuai dengan kebutuhan spesifik, seperti prioritas pada recall atau precision.

Untuk pengembangan lebih lanjut, disarankan untuk memperluas dataset dengan berita dari sumber yang lebih beragam agar model dapat menangani variasi bahasa dan konteks yang lebih luas. Studi selanjutnya juga dapat mengeksplorasi algoritma tambahan seperti Gradient Boosting Machines atau deep learning, seperti LSTM atau Transformer, untuk meningkatkan performa deteksi hoax, khususnya pada dataset yang sangat besar dan kompleks. Penelitian juga dapat memperluas fokus pada analisis semantik melalui representasi teks berbasis word embeddings seperti Word2Vec atau FastText untuk menangkap hubungan antar kata yang lebih dalam. Langkah ini dapat menghasilkan sistem deteksi hoax yang lebih adaptif dan akurat dalam menghadapi tantangan penyebaran informasi palsu di era digital.

DAFTAR PUSTAKA

[1] C. B. Devina, D. C. Iswari, G. C. B. Goni, dan D. K. Lirungan, "Tinjauan Hukum Kriminalisasi Berita Hoax: Menjaga Persatuan vs. Kebebasan Berpendapat," *Kosmik Huk.*, vol. 21, no. 1, hlm. 44, Feb 2021, doi: 10.30595/kosmikhukum.v21i1.8874.

[2] J. E. Latupeirissa, J. D. Pasalbessy, E. Z. Leasa, dan C. Tuhumury, "Penyebaran Berita Bohong (HOAX) Pada Masa Pandemi Covid-19 dan Upaya Penanggulangannya di Provinsi Maluku," *J. BELO*, vol. 6, no. 2, hlm. 179–194, Feb 2021, doi: 10.30598/belovol6issue2page179-194.

[3] L. M. W. Pangestika *dkk.*, "Identifikasi Potensi Desa dan Kebutuhan Pengajaran Anti Hoax (Studi Kasus Desa Pucanganom, DIY)," *J. Atma Inovasia*, vol. 1, no. 1, Art. no. 1, Jan 2021, doi: 10.24002/jai.v1i1.3915.

[4] F. Farhan, R. Aziz, dan I. Nurdin, "Penggunaan media sosial selama pandemi COVID-19 dan dampaknya terhadap penyebaran hoax," *J. Media Sos. Dan Inf.*, vol. 7, no. 3, hlm. 95–110, 2022.

[5] E. Susanti dan L. Nurmiati, "Pengaruh literasi digital terhadap kemampuan masyarakat dalam menyaring informasi hoax," *J. Teknol. Inf. Dan Komun.*, vol. 5, no. 1, hlm. 88–97, 2022.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

[6] H. Putra dan M. Patra, "Pengaruh hoax terhadap persepsi politik selama pemilu di Indonesia," *J. Polit. Dan Demokr.*, vol. 12, no. 3, hlm. 187–202, 2023.

[7] D. Dharmansyah, H. Arifin, dan A. Suryadi, "Dampak psikologis dari informasi hoax mengenai vaksinasi COVID-19," *J. Psikol. Kesehat.*, vol. 10, no. 1, hlm. 123–140, 2023.

[8] Naseer, "Sistem prediksi berita palsu tentang virus COVID-19 menggunakan algoritma support vector machine (SVM)," *Naratif J. Nas. Ris. Apl. Dan Tek. Inform.*, 2023, doi: 10.53580/naratif.v5i1.187.

[9] P. Pratiwi, "Penggunaan model klasifikasi bahasa Indonesia untuk deteksi hoax," *J. Pemrosesan Bhs. Alami*, vol. 5, no. 4, hlm. 123–138, 2022.

[10] Roshinta, "Sistem deteksi berita hoax berbahasa Indonesia bidang kesehatan," *Remik Ris. Dan E-J. Manaj. Inform. Komput.*, 2023, doi: 10.33395/remik.v7i2.12369.

[11] R. Rifai, "Model hibrida untuk deteksi berita hoax dengan akurasi tinggi," *J. Sist. Inf.*, vol. 10, no. 3, hlm. 90–105, 2023.

[12] I. M. K. Karo, "Hoax Detection on Indonesian Tweets Using Naïve Bayes Classifier With TF-IDF," *J. Inf. Syst. Res. Josh*, vol. 4, no. 3, hlm. 914–919, 2023, doi: 10.47065/josh.v4i3.3317.

[13] L. Holla, "An Improved Fake News Detection Model Using Hybrid Time Frequency-Inverse Document Frequency for Feature Extraction and AdaBoost Ensemble Model as a Classifier," *J. Adv. Inf. Technol.*, vol. 15, no. 2, hlm. 202–211, 2024, doi: 10.12720/jait.15.2.202-211.

[14] D. P. Putra dan E. B. Setiawan, "Hoax Detection Using Long Short-Term Memory (LSTM) and Gate Recurrent Unit (GRU) on Social Media," *Build. Inform. Technol. Sci. Bits*, vol. 4, no. 4, 2023, doi: 10.47065/bits.v4i4.3084.

[15] H. B. Aji, "Detecting Hoax Content on Social Media Using Bi-LSTM and RNN," *Build. Inform. Technol. Sci. Bits*, vol. 5, no. 1, 2023, doi: 10.47065/bits.v5i1.3585.

[16] F. R. Tama, "Fake News (Hoaxes) Detection on Twitter Social Media Content Through Convolutional Neural Network (CNN) Method," *Jinav J. Inf. Vis.*, vol. 4, no. 1, hlm. 70–78, 2023, doi: 10.35877/454ri.jinav1525.

[17] G. B. Firmanesha, "Detecting Hoax News Regarding the Covid-19 Vaccine Using Levenshtein Distance," *J. Bumigora Inf. Technol. Bite*, vol. 4, no. 2, hlm. 133–142, 2022, doi: 10.30812/bite.v4i2.2023.

[18] A. Y. Prayoga, A. I. Hadiana, dan F. R. Umbara, "Deteksi Hoax Pada Berita Online Bahasa Inggris Menggunakan Bernoulli Naïve Bayes Dengan Ekstraksi Fitur Tf-Idf," *J. Syntax Admiration*, vol. 2, no. 10, hlm. 1808–1823, 2021, doi: 10.46799/jsa.v2i10.327.

[19] I. Maulita dan A. Wahid, "Prediksi Magnitudo Gempa Menggunakan Random Forest, Support Vector Regression, XGBoost, LightGBM, dan Multi-Layer Perceptron Berdasarkan Data Kedalaman dan Geolokasi (Predicting Earthquake Magnitude Using Random Forest, Support Vector Regression, XGBoost, LightGBM, and Multi-Layer Perceptron Based on Depth and Geolocation Data)," *J. Pendidik. Dan Teknol. Indones.*, vol. 4, hlm. 221–232, Mei 2024, doi: 10.52436/1.jpti.470.

[20] A. M. Wahid, L. Afuan, dan F. S. Utomo, "ENHANCING COLLABORATION DATA MANAGEMENT THROUGH DATA WAREHOUSE DESIGN: MEETING BAN-PT ACCREDITATION AND KERMA REPORTING REQUIREMENTS IN HIGHER EDUCATION," *J. Tek. Inform. Jutif*, vol. 5, no. 6, Art. no. 6, Des 2024, doi: 10.52436/1.jutif.2024.5.6.1747.

[21] A. M. Wahid, T. Hariguna, dan G. Karyono, "Optimizing Feature Extraction for Website Visuals: A Comparative Study of AlexNet and Inception V3," dalam *2024 12th International Conference on Cyber and IT Service Management (CITSM)*, Okt 2024, hlm. 1–6. doi: 10.1109/CITSM64103.2024.10775681.

[22] A. D. Riyanto, A. M. Wahid, dan A. A. Pratiwi, "ANALYSIS OF FACTORS DETERMINING STUDENT SATISFACTION USING DECISION TREE, RANDOM FOREST, SVM, AND NEURAL NETWORKS: A COMPARATIVE STUDY," *J. Tek. Inform. Jutif*, vol. 5, no. 4, Art. no. 4, Jul 2024, doi: 10.52436/1.jutif.2024.5.4.2188.

[23] Berlilana dan A. M. Wahid, "Time Series Analysis of Bitcoin Prices Using ARIMA and LSTM for Trend Prediction," *J. Digit. Mark. Digit. Curr.*, vol. 1, no. 1, Art. no. 1, Mei 2024, doi: 10.47738/jdmdc.v1i1.1.

[24] B. Berlilana, A. M. Wahid, D. Fortuna, A. N. A. Saputra, dan G. Bagaskoro, "Exploring the Impact of Discount Strategies on Consumer Ratings: An Analytical Study of Amazon Product Reviews," *J. Appl. Data Sci.*, vol. 5, no. 1, Art. no. 1, Jan 2024, doi: 10.47738/jads.v5i1.163.