E-ISSN: 3031-1640
DOI: doi-No.
Received: 17/01/2025, Revised: 25/01/2025, Published: 29/01/2025

PERBANDINGAN KLASIFIKASI BERITA HOAX POLITIK PADA MEDIA SOSIAL X MENGGUNAKAN ALGORITMA NAÏVE BAYES DAN RANDOM FOREST

Rika Astuti[1] , Angel Tiarma Sipahutar[2]
[1]Teknologi Informasi ,Universitas Siber Indonesia , rika.astuti93@gmail.com
[2]Teknologi Informasi, Universitas Siber Indonesia , 13210036@cyber-univ.ac.id

Korespondensi: Rika Astuti [1]

Abstract: *Social media has become a major source of information as well as a fertile ground for the spread of hoax news, particularly political news. This study examines two classification algorithms, Naive Bayes and Random Forest, to detect hoax news. The data used consists of political news labeled as hoax or non-hoax. The results show that the Random Forest algorithm has better accuracy and performance compared to Naive Bayes in terms of precision, recall, and F1-score. In conclusion, Random Forest is more effective for classifying political hoax news on social media. This research provides important insights into the application of machine learning algorithms in text classification and emphasizes the importance of selecting the right algorithm for optimal results.*

Keywords: *Hoax News Classification, Naive Bayes Algorithm, Random Forest Algorithm, Social Media, Political News, Machine Learning.*

Abstract: Media sosial X telah menjadi sumber utama informasi sekaligus tempat penyebaran berita hoax, khususnya berita politik. Penelitian ini menguji dua algoritma klasifikasi, Naive Bayes dan Random Forest, untuk mendeteksi berita hoax pada platform tersebut. Data yang digunakan adalah berita politik yang telah diberi label hoax atau non-hoax. Hasil penelitian menunjukkan bahwa algoritma Random Forest memiliki akurasi dan performa yang lebih baik dibandingkan dengan Naive Bayes dalam hal presisi, recall, dan F1-score. Hasil penelitian menunjukkan bahwa Random Forest lebih efektif untuk klasifikasi berita hoax politik di media sosial X. Penelitian ini memberikan wawasan penting tentang penerapan algoritma machine learning dalam klasifikasi teks dan menekankan pentingnya pemilihan algoritma yang tepat untuk hasil yang optimal.

Kata Kunci: Klasifikasi Berita *Hoax,* Algoritma *Naïve Bayes,* Algoritma *Random Forest,* Media Sosial X, Berita Politik, *Machine Learning*.

PENDAHULUAN

Seiring dengan perkembangan teknologi informasi yang semakin maju dengan sangat pesat pada saat ini, informasi dapat ditemukan kapan dan dimana saja. masyarakat bisa mendapatkan berbagai informasi baik dari mulut ke mulut, maupun melalui media sosial. Banyak orang yang menjadikan media sosial sebagai tempat untuk memperoleh informasi. Namun, kemudahan dan kecepatan ini juga membawa konsekuensi, yaitu maraknya penyebaran informasi yang salah dan menyesatkan, atau yang dikenal dengan hoax (Atmojo & Pratiwi, n.d., 2021).

Dua algoritma klasifikasi yang umum digunakan untuk mendeteksi hoax di media sosial X adalah Naive Bayes dan Random Forest. Algoritma Naive Bayes adalah algoritma probabilistik yang sederhana dan mudah diimplementasikan (Putri et al., 2022). Algoritma ini bekerja dengan menghitung probabilitas kemunculan dari kata- kata negatif atau fitur tertentu dalam sebuah tweet, dan (mencari nilai akurasi dari kedua metode tsb). Penggunaan algoritma Random Forest dalam pengklasifikasian berita hoax di media sosal X menawarkan sejumlah keunggulan yang signifikan. Pertama-tama, media sosial X adalah platform yang dinamis dengan volume data yang besar. Algoritma Random Forest secara efisien dapat menangani dataset yang luas dan kompleks, memproses informasi dari banyak fitur dengan baik. Hal ini memungkinkan model untuk menghadapi tantangan dari volume tweet yang terus meningkat tanpa mengorbankan kualitas atau keakuratan dalam pengklasifikasian.Diharapkan bahwa penelitian ini akan memberikan kontribusi yang berarti dalam upaya melawan penyebaran berita hoax di X dan platform media sosial lain khususnya berita pada bidang politik. Dengan menggunakan algoritma Random Forest dalam pengklasifikasian berita hoax, diharapkan kita dapat lebih efektif dalam mendeteksi dan mengidentifikasi konten yang tidak benar atau menyesatkan (Saputra, E. N., & Yuliana, M. E. 2023). Hal ini akan membantu dalam melindungi masyarakat dari dampak negatif hoax, seperti keresahan sosial, gangguan aktivitas, dan bahkan risiko perpecahan antara masyarakat akibat berita politik yang tidak benar. Dengan pengembangan metode yang efektif untuk mengatasi masalah ini, diharapkan dapat diciptakan lingkungan daring yang lebih aman dan terpercaya, di mana informasi politik yang disajikan dapat diverifikasi dan dipercaya. Dengan demikian, penelitian ini tidak hanya berpotensi meningkatkan keamanan dan stabilitas platform media sosial, tetapi juga memberikan kontribusi positif terhadap kesejahteraan dan keamanan masyarakat secara keseluruhan.

TINJAUAN LITERATUR

Hoax adalah berita bohong yang sengaja dibuat yang mengaburkan atau memutarbalikkan fakta untuk tujuan tertentu (Pertanggungjawaban Pidana Penyebaran Berita Hoax di Media Sosial. 2023). Hoax adalah suatu kata yang digunakan untuk menunjukkan pemberitaan palsu atau usaha untuk menipu atau mengakali pembaca/pendengarnya untuk mempercayai sesuatu yang biasanya digunakan dalam media sosial (Mauludi, S. 2018). Hoax adalah kabar, informasi, berita palsu, atau bohong. Sedangkan dalam KBBI disebut dengan hoaks yang artinya berita bohong (Wijayanti, 2019). Hoax menjadi semakin relevan di era digital, di mana informasi dapat dengan mudah dan cepat tersebar melalui media sosial, aplikasi pesan singkat, dan berbagai platform daring lainnya. Hoax dapat berbentuk teks, gambar, video, atau kombinasi dari semuanya, dan sering kali disamarkan sebagai berita atau informasi yang sahih untuk mengelabui penerima.

Menurut (Lestari, et al. (2022), klasifikasi adalah menentukan sebuah record data baru ke salah satu dari beberapa kategori yang telah didefinisikan sebelumnya. Klasifikasi merupakan peran dalam data mining yang menggunakan metode pendekatan prediktif. Secara umum, proses klasifikasi dapat dilakukan dalam dua tahap, yaitu proses belajar dari data pelatihan dan klasifikasi kasus. Pada proses belajar, algoritma klasifikasi mengolah data training untuk menghasilkan sebuah model. Setelah model diuji dan dapat diterima, pada tahap klasifikasi model tersebut digunakan untuk memprediksi kelas dari kasus baru untuk membantu proses pengambilan keputusan.

Naive Bayes merupakan algoritma data mining yang dapat digunakan untuk melakukan klasifikasi dan merupakan algoritma yang akan memberikan hasil kemungkinan yang akan terjadi di masa akan datang dari hasil klasifikasi (Gunawan et al., 2021). Naive Bayes mengasumsikan bahwa setiap fitur dalam data adalah independen satu sama lain, meskipun dalam dunia nyata, ini mungkin tidak selalu benar. Asumsi ini membuat Naive Bayes sederhana dan cepat dalam pelatihan dan prediksi, karena model hanya perlu menghitung probabilitas munculnya setiap fitur terhadap setiap kelas.

Random Forest adalah salah satu metode data mining untuk menentukan klasifikasi. Algoritma Random Forest sendiri berupa pohon klasifikasi dengan jumlah koneksi yang banyak untuk mencari akurasi yang maksimal (Wantoro et al., 2021). Random Forest adalah kombinasi dari masing-masing teknik pohon keputusan yang ada, lalu kemudian digabung dan dikombinasikan ke dalam suatu model. Random Forest menggabungkan kekuatan dari banyak model pembelajaran untuk menciptakan prediksi yang lebih kuat dan stabil.

METODE

1) Pendekatan Penelitian

Menurut (Arikunto, Suharsimi.2019) Metode penelitian adalah cara utama yang digunakan peneliti untuk mencapai tujuan dan menentukan jawaban atas masalah yang diajukan. Metode yang digunakan pada penelitian ini adalah eksperimen.

2) Fokus Penelitian

Berdasarkan permasalahannya, penulis membatasi untuk fokus penelitian pada berita politik pada media sosial X yang akan di bagi menjadi dua yaitu hoax dan valid dan kinerja algoritma akan di evaluasi berdasarkan metrik accuracy, presisi, recall, dan F1-Score dalam mengklasifikasikan berita dengan tepat.

3) Teknik Pengumpulan Data

Teknik pengumpulan data dapat diartikan sebagai rosedural komponen sistematis yang saling berkaitan dalam mencapai tujuan yang di tetapkan (Nasarudin, 2023). Teknik pengumpulan data dalam penelitian ini dilakukan secara daring dengan memanfaatkan platform media sosial X sebagai sumber utama. Media sosial X dipilih karena memiliki pengguna aktif yang sangat banyak dan sering menjadi tempat penyebaran berbagai jenis informasi, termasuk berita hoax. Pengumpulan data dilakukan dengan beberapa langkah berikut:

1. Penelitian ini akan menggunakan dua kategori utama sumber data. Data valid akan diperoleh dari sumber berita terkemuka dan terpercaya seperti CNN, Kompas, dan Tempo.

Sumber berita ini dikenal dengan standar jurnalistik yang tinggi dan integritas informasi. Sementara itu, data hoax akan dikumpulkan dari *Turn Back Hoax* yang fokus pada mengidentifikasi dan membantah berita palsu.

2. Setelah data dikumpulkan, narasi dari berita yang diperoleh akan dianalisis untuk menentukan topik dan *keyword* yang relevan. Proses ini melibatkan ekstraksi tema utama dan istilah penting dari berita *valid* dan *hoax*, yang akan menjadi dasar untuk langkah selanjutnya.Setiap artikel atau postingan yang dikumpulkan diberi label apakah mengandung informasi hoax atau bukan berdasarkan verifikasi dari sumber-sumber terpercaya. Label ini digunakan sebagai indikator dalam kolom hoax pada dataset, yang membantu dalam proses pelatihan dan evaluasi model pembelajaran mesin yang digunakan dalam penelitian ini.
3. Berdasarkan *keyword* yang telah diidentifikasi, tahap berikutnya adalah melakukan *scraping* data dari platform X (X.com). Teknik *scraping* ini bertujuan untuk mengumpulkan berita terkait dengan *keyword* yang telah ditentukan, yang akan memperluas kumpulan data dan memungkinkan analisis yang lebih mendalam.

4) Teknik Analisis Data

Dalam penelitian ini, metode CRISP-DM (Cross-Industry Standard Process for Data Mining) digunakan bersama dengan algoritma Naive Bayes dan Random Forest. Setelah data dikumpulkan, langkah selanjutnya adalah melakukan analisis terhadap data tersebut. Analisis data melibatkan beberapa tahap, termasuk :

1. Labelling data

Pada tahapan ini, data yang sudah didapatkan dari proses *scraping* akan di berikan label sesuai dengan kata kunci. Dimana apabila data tersebut merupakan data hasil scraping dari *keyword* sumber data *valid* (CNN, Kompas, Tempo) maka hasil scraping tersebut akan dilabelkan dengan angka 1. Dan sebaliknya, apabila data yang merupakan hasil scraping dari keyword sumber data paltform *hoax* (*Turn Back Hoax*) akan diberikan label 0.

2. *Preprocessing*

Pada tahap ini, data yang telah dikumpulkan akan melakukan proses preprocessing untuk membersihkan data sebelum di olah atau dilakukan analisis lebih lanjut. Tahapan preprocessing yang akann dilakukan adalah *tokenizing*, *cleaning*, *case folding*, *stemming*, *stopwords removal*

3. Pembagian Dataset

Setelah melalui tahapan preprocessing data akan di bagi menjadi dua bagian yaitu data training dan data testing dengn skala 80:20. Dimana data training (80%) dan data testing (20%)

4. Ekstraksi *Fitur*

Setelah data dibagi, tahap selanjtnya adalah ekstraksi fitur dari data yang sudah didapatkan menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). TF-IDF akan menghasilkan representasi numerik dari setiap data yang akan digunakan sebagai input untuk algoritma machine learning.

5. *Modelling*

Tahapan ini melibatkan Kedua algoritma pada penelitian ini yaitu *Naive Bayes* dan *Random Forest* akan dilatih menggunakan data training

6. Evaluasi (Evaluation)

Menggunakan metrik evaluasi Akurasi (Proporsi yang benar dari total prediksi), Presisi (Proporsi *true postitve* dari semua yang diprediksi positif), *Recall* (*Proporsi true positive* dari semua yang seharusnya positif), F1-*Score* (*Harmonic mean* dari presisi dan *recall*) dan *Confusion matrix* Untuk memahami dimana model melakukan kesalahan klasifikasi, misalnya memprediksi berita *hoax* sebagai *valid* atau sebaliknya.

7. Analisis dan Interpretasi hasil

Tahapan ini adalah membandingkan kinerja setiap model algoritma dan menentukan mana yang memberikan hasil terbaik dalam mengklasifikasikan berita hoax politik pada media sosial X
Pada tahap ini merupakan proses yang akan digunakan dalam mengimplementasikan teknik

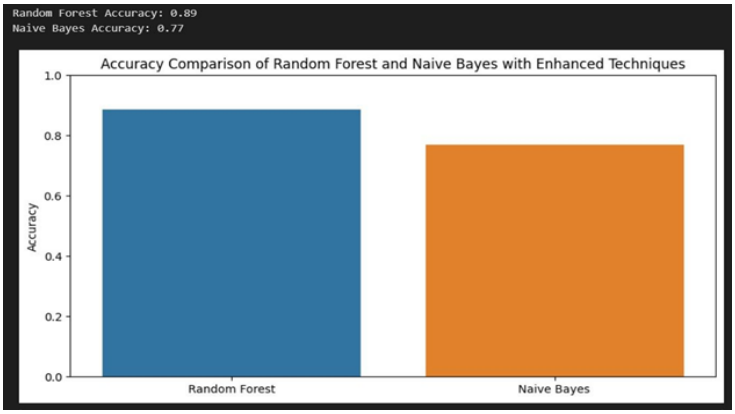
HASIL DAN PEMBAHASAN

Visualisasi Perbandingan Kedua Model

```
plt.figure(figsize=(10, 5))
accuracy_scores = [rf_accuracy, nb_accuracy]
classifiers = ['Random Forest', 'Naive Bayes']
sns.barplot(x=classifiers, y=accuracy_scores)
plt.ylim(0, 1)
plt.title('Accuracy Comparison of Random Forest and Naive Bayes with Enhanced Techniques')
plt.ylabel('Accuracy')
plt.show()
```

Gambar 1. Code Untuk Membuat Grafik Bar Perbandingan Kedua Model

Gambar 1. tersebut menunjukkan code yang digunakan untuk membuat grafik batang yang membandingkan akurasi model Random Forest dan Naive Bayes. code tersebut mendefinisikan daftar skor accuracy **(`rf_accuracy` dan `nb_accuracy`)** dan nama model (`**Random Forest**` dan `**Naive Bayes'**). Grafik batang dibuat menggunakan `seaborn.barplot` dengan nama model pada sumbu x dan skor akurasi pada sumbu y. Batas sumbu y diatur dari 0 hingga 1 untuk mencerminkan rentang *accurracy.* Judul dan label sumbu y ditambahkan untuk memperjelas bahwa grafik ini membandingkan akurasi kedua model dengan teknik yang ditingkatkan. Dan, grafik ditampilkan dengan **code `plt.show()`.**



Gambar 2. *Visualisasi Grafik Perbandingan Kedua Model*

Gambar 2 di atas merupakan hasil grafik batang menunjukkan bahwa model Random Forest memiliki akurasi yang lebih tinggi dibandingkan dengan model Naive Bayes. Secara spesifik, Random Forest mencapai accuracy 0.89, sedangkan Naive Bayes hanya mencapai accuracy 0.77. Ini menunjukkan bahwa Random Forest lebih baik dalam memprediksi data dengan benar dibandingkan Naive Bayes dalam konteks dataset dan fitur yang digunakan. Perbedaan ini mungkin disebabkan oleh kemampuan Random Forest untuk menangani kompleksitas data dengan membangun banyak pohon keputusan dan mengambil suara mayoritas, sementara Naive Bayes membuat asumsi independensi yang mungkin tidak sesuai dengan karakteristik data.

KESIMPULAN

Berdasarkan hasil penelitian mengenai klasifikasi berita *hoax* politik pada media sosial X dengan menggunakan dua algoritma *machine learning*, yaitu Naïve Bayes dan Random Forest, diperoleh hasil sebagai berikut:

1. Evaluasi Model Naïve Bayes: Model ini menunjukkan nilai akurasi sebesar 77%, presisi sebesar 78%, recall sebesar 77%, dan F1-Score sebesar 77%.
2. Evaluasi Model Random Forest: Model ini menunjukkan nilai akurasi sebesar 89%, presisi sebesar 83%, recall sebesar 89%, dan F1-Score sebesar 89%.

Random Forest menunjukkan akurasi yang lebih tinggi dalam menangani dataset yang besar dan kompleks. Proses pengklasifikasian mengikuti tahapan metode CRISP-DM (*Cross-Industry Standard Process for Data Mining*), yang meliputi pemahaman data, persiapan data, permodelan, dan evaluasi membantu memastikan penelitian dilakukan dengan sistematis dan terstruktur. Pentingnya pemahaman dan persiapan data yang baik diungkapkan sebagai faktor kunci dalam membangun model yang akurat. Selain itu, penelitian ini menekankan peran penting teknologi dalam mendeteksi hoax, menunjukkan bahwa algoritma machine learning dapat digunakan secara efektif untuk mengurangi penyebaran informasi yang salah di media sosial. Hasil dari penelitian ini diharapkan dapat berkontribusi pada upaya melawan hoax dan menciptakan lingkungan daring yang lebih aman dan terpercaya

DAFTAR PUSTAKA

Atmojo, M. E., Darumurti, A., & Pratiwi, V. P. (2021) Application of Position Auction- Based Merit System in Bantul Regency in 2018. JPPUMA: Jurnal Ilmu Pemerintahan dan Sosial Politik UMA (Journal of Governance and Political Social UMA), 9(2), 129-138.

Gunawan, M., Zarlis, M., & Roslina, R. (2021). Analisis Komparasi Algoritma Naïve Bayes dan K-Nearest Neighbor Untuk Memprediksi Kelulusan Mahasiswa.

Lestari, S., Akmaludin, A., & Badrul, M. (2020) Implementası Klasifikasi Naive Bayes Untuk Prediksi Kelayakan Pemberian Pinjaman Pada Koperasi Amugerah Bintang Cemerlang. PROSISKO: Jurnal Pengembangan Riset dan Observasi Sistem Komputer, 7

Mc Kinney, W ,(2020) , Pyton for Data Analysis (2nd ed) . O' Really Media.

Mauludi, S. (2018). Awas hoax! cerdas menghadapi pencemaran nama baik, ujaran kebencian & hoax. Indonesia PT Elex Media Komputindo.

Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E., & Zanini, F. (2022). Analysing high- throughput sequencing data in Python with HTSeq 2.0. Bioinformatics, 38(10). 2943-2945.

Saputra, E. N., & Yuliana, M. E. (2023). Mitigasi Hoax Di Era Disrupsi Melalui Literasi Digital Hate Speech Pada Pengguna Layanan Internet, JPNM Jurnal Pustaka Nusantara Multidisiplin, 1(4).

Wijayanti, S (2019) Art Kara Hoar Kama Besar Bahasa Indonesia (KBBI).